

A. Query Input and database selection

Accepted Input Formats

Query sequence(s) to be used for a BLAST search should be pasted in the **'Search'** text area. It accepts a number of different types of input and automatically determines the format or the input. To allow this feature there are certain conventions required with regard to the input of identifiers (e.g., accessions or gi's). These are described in 3) below. Accepted input types are FASTA, bare sequence, or sequence identifiers .

1. FASTA

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (define) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. It is recommended that all lines of text be shorter than 80 characters in length. An example sequence in FASTA format is:

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVVLVNAIYFKGMWKTAFNAEDTREMPPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPPFASGDLMSMLVLLPDEVSDLERIEKTINFEKLTWWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPSESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Blank lines are not allowed in the middle of FASTA input.

Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters (see below). Before submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue). The nucleic acid codes supported are:

A	adenosine	C	cytidine	G	guanine
T	thymidine	N	A/G/C/T (any)	U	uridine
K	G/T (keto)	S	G/C (strong)	Y	T/C (pyrimidine)
M	A/C (amino)	W	A/T (weak)	R	G/A (purine)
B	G/T/C	D	G/A/T	H	A/C/T
V	G/C/A	-	gap of undetermined length		

NOTE:

¹The degenerate nucleotide codes in red are treated as mismatches in nucleotide alignment. Too many such degenerate codes within an input nucleotide query will cause *blast.cgi* to reject the input. For protein queries, too many nucleotide-like code (A,C,G,T,N) may also cause similar rejection.

²*blast.cgi* will not take "-" in the query. To represent gaps, use a string of N's instead.

For those programs that use amino acid query sequences (BLASTP and TBLASTN), the accepted amino acid codes are:

A	alanine	O	pyrrolysine
B	aspartate/asparagine	P	proline
C	cystine	Q	glutamine
D	aspartate	R	arginine
E	glutamate	S	serine
F	phenylalanine	T	threonine
G	glycine	U	selenocysteine
H	histidine	V	valine
I	isoleucine	W	tryptophan
J	*leucine	Y	tyrosine
K	lysine	Z	glutamate/glutamine
L	leucine	X	any
M	methionine	*	translation stop

example to limit matches to the region from 24 to 200 of a query sequence, you would enter 24 in the "From" field and 200 in the "To" field. If one of the limits you enter is out of range, the intersection of the [From,To] and [1,length] intervals will be searched, where length is the length of the whole query sequence.

Databases available for BLAST search

The BLAST pages offer several different databases for searching. Some of these, like SwissProt and PDB are compiled outside of NCBI. Other like ecoli, dbEST and month, are subsets of the NCBI databases. Other "virtual Databases" can be created using the ["Limit by Entrez Query"](#) option.

• Peptide Sequence Databases

- **nr**
All non-redundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF
- **refseq**
RefSeq protein sequences from NCBI's Reference Sequence Project.
- **swissprot**
Last major release of the SWISS-PROT protein sequence database (no updates).
- **pat**
Proteins from the Patent division of GenPept.
- **pdb**
Sequences derived from the 3-dimensional structure from [Brookhaven Protein Data Bank](#).
- **month**
All new or revised GenBank CDS translation+PDB+SwissProt+PIR+PRF released in the last 30 days.
- **env_nr**
Protein sequences from environmental samples.

• Nucleotide Sequence Databases

- **Human genomic plus transcript** (default)
Current refseq and alternative human genome assemblies and transcripts, with link to MapView.
- **Mouse genomic plus transcrip**
Current refseq and alternative mouse genome assemblies and transcripts, with link to MapView.
- **nr**
All GenBank + RefSeq Nucleotides + EMBL + DDBJ + PDB sequences (excluding HTGS0,1,2, EST, GSS, STS, PAT, WGS, and env_nt). No longer "non-redundant".
- **refseq_rna**
RNA entries from NCBI's Reference Sequence project
- **refseq_genomic**
Partial genomic entries from NCBI's Reference Sequence project
- **est**
Database of GenBank + EMBL + DDBJ sequences from [EST Divisions](#)
- **est_human**
Human subset of est.
- **est_mouse**
Mouse subset.
- **est_others**
Non-Mouse, non-Human subset of est.
- **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 from [Brookhaven Protein Data Bank](#)
- **month**
All new or revised GenBank + EMBL + DDBJ + PDB sequences released in the last 30 days.
- **dbsts**
Database of GenBank+EMBL+DDBJ sequences from STS Divisions .
- **chromosome**
A database with complete genomes and chromosomes from the [NCBI Reference Sequence project](#).
- **wgs**
A database for whole genome shotgun sequence entries.
- **env_nt**
Nucleotide sequences from environmental samples, including those from Sargasso Sea and Mine Drainage projects.

Return alignment endpoints only

This is the simplest BLAST output format, only available for megablast. Selection of this will disable future reformatting of the same BLAST search through the search RID.

Hits computed

It is possible to speed up search by specifying maximum number of hits to be computed. This is only available to Trace megablast page.

CDD Search

This function is relevant only to Protein BLAST pages. When activated, it will search the input protein sequences against the Conserved Domain Database (CDD) at the same time. Conserved domains matching to the query may provide additional insight into the possible function of the query.

CDD is a database containing a collection of protein alignment profiles derived from two outside collections, [Smart](#) and [Pfam](#), plus entries created within NCBI: COG and cd. For more information please see the [CDD homepage](#).

Choose a translation

This is only relevant "Translated" BLAST pages. The translations include:

- **blastx**

which compares a nucleotide query sequence translated in all reading frames against a protein sequence database

- **tblastn**

which compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.

- **tblastx**

which compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Due its high computation cost and background noise, tblastx should only be used as a last resort. Users with regular and large scale needs of tblastx should install command line BLAST and run the searches locally.

Query Genetic Code

Genetic code to be used in blastx and tblastx translation of the query. See list of Genetic Codes in [Taxonomy](#).

B. BLAST Search Parameters

Limit by Entrez Query

A BLAST search can be limited to the result of an Entrez query against the database chosen. This restricts the search to a subset of entries from that database fitting the requirement of the Entrez query. Terms normally accepted by Entrez nucleotide or protein searches are accepted here. Examples are given below.

- **protease NOT hiv1[organism]**

This will limit a BLAST search to all proteases, except those in HIV 1.

- **1000:2000[slen]**

This limits the search to entries with lengths between 1000 to 2000 bases for nucleotide entries, or 1000 to 2000 residues for protein entries.

- **Mus musculus[organism] AND biomol_mrna[properties]**

This limits the search to mouse mRNA entries in the database. For common organisms, one can also select from the pulldown menu.

- **10000:100000[mlwt]**

This is yet another example usage, which limits the search to protein sequences with calculated molecular weight between 10 kD to 100 kD.

For help in constructing Entrez queries please see the "[Writing Advanced Search Statements](#)" section of the Entrez Help document. Knowing the content of a database and applying the Entrez terms accordingly are important. For example, biomol_mrna[prop] should not be applied to htgs or chromosome database since they do not contain mRNA entries!

Compositional adjustments

Amino acid substitution matrices may be adjusted in various ways to compensate for the amino acid compositions of the sequences being compared. The simplest adjustment is to scale all substitution scores by an analytically determined constant, while leaving the gap scores fixed; this procedure is called "composition-based statistics" (Schaffer et al., 2001). The resulting scaled scores yield more accurate E-values than standard, unscaled scores. A more sophisticated approach adjusts each score in a standard

substitution matrix separately to compensate for the compositions of the two sequences being compared (Yu et al., 2003; Yu and Altschul, 2005; Altschul et al., 2005). Such "compositional score matrix adjustment" may be invoked only under certain specific conditions for which it has been empirically determined to be beneficial (Altschul et al., 2005); under all other conditions, composition-based statistics are used. Alternatively, compositional adjustment may be invoked universally.

- [1] Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001) "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Res.* 29:2994-3005.
- [2] Yu, Y.-K., Wootton, J.C. and Altschul, S.F. (2003) "The compositional adjustment of amino acid substitution matrices," *Proc. Natl. Acad. Sci. USA* 100:15688-15693.
- [3] Yu, Y.-K. and Altschul, S.F. (2005) "The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions," *Bioinformatics* 21:902-911.
- [4] Altschul, S.F., Wootton, J.C., Gertz, E.M., Agarwala, R., Morgulis, A., Schaffer, A.A. and Yu, Y.-K. (2005) "Protein database searches using compositionally adjusted substitution matrices," *FEBS J* 272(20):5101-9.

Filter

• Filter (Low-complexity)

This function mask off segments of the query sequence that have low compositional complexity, as determined by the [SEG](#) program of Wootton and Federhen (Computers and Chemistry, 1993) or, for BLASTN, by the [DUST](#) program of Tatusov and Lipman. Filtering can eliminate statistically significant but biologically uninteresting reports from the blast output (e.g., hits against common acidic-, basic- or proline-rich regions), leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences.

Filtering is only applied to the query sequence (or its translation products), not to database sequences. Default filtering is DUST for BLASTN, SEG for other programs.

It is not unusual for nothing at all to be masked by SEG, when applied to sequences in SWISS-PROT or refseq, so filtering should not be expected to always yield an effect. Furthermore, in some cases, sequences are masked in their entirety, indicating that the statistical significance of any matches reported against the unfiltered query sequence should be suspect. This will also lead to search error when default setting is used.

• Filter (Human repeats)

This option masks Human repeats (LINE's, SINE's, plus retroviral repeats) and is useful for human sequences that may contain these repeats. Filtering for repeats can increase the speed of a search especially with very long sequences (>100 kb) and against databases which contain large number of repeats (htgs). This filter should be checked for genomic queries to prevent potential problems that may arise from the numerous and often spurious matches to those repeat elements.

For more information please see ["Why does my search timeout on the BLAST servers?"](#) in the BLAST Frequently Asked Questions.

• Filter (Mask for lookup table only)

BLAST searches consist of two phases, finding hits based upon a lookup table and then extending them. This option masks only for purposes of constructing the lookup table used by BLAST so that no hits are found based upon low-complexity sequence or repeats (if repeat filter is checked). The

BLAST extensions are performed without masking and so they can be extended through low-complexity sequence.

- **Mask Lower Case**

With this option selected you can cut and paste a FASTA sequence in upper case characters and denote areas you would like filtered with lower case. This allows you to customize what is filtered from the sequence during the comparison to the BLAST databases.

One can use different combinations of the above filter options to achieve optimal search result.

Word-size

BLAST is a heuristic that works by finding word-matches between the query and database sequences. One may think of this process as finding "hot-spots" that BLAST can then use to initiate extensions that might eventually lead to full-blown alignments. For nucleotide-nucleotide searches (i.e., "blastn") an exact match of the entire word is required before an extension is initiated, so that one normally regulates the sensitivity and speed of the search by increasing or decreasing the word-size. For other BLAST searches non-exact word matches are taken into account based upon the similarity between words. The amount of similarity can be varied so one normally uses just the word-sizes 2 and 3 for these searches.

Expect

This setting specifies the statistical significance threshold for reporting matches against database sequences. The default value (10) means that 10 such matches are expected to be found merely by chance, according to the stochastic model of Karlin and Altschul (1990). If the statistical significance ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported.

Reward and Penalty for Nucleotide Programs

Many nucleotide searches use a simple scoring system that consists of a "reward" for a match and a "penalty" for a mismatch. The (absolute) reward/penalty ratio should be increased as one looks at more divergent sequences. A ratio of 0.33 (1/-3) is appropriate for sequences that are about 99% conserved; a ratio of 0.5 (1/-2) is best for sequences that are 95% conserved; a ratio of about one (1/-1) is best for sequences that are 75% conserved [1].

To ensure BLAST returns more reliable statistics for blastn searches, NCBI has put new restraint on the allowed reward/penalty pairs and their associated gap existence and gap extension penalties. See [News on blast 2.2.13](#) for more information.

[1] States DJ, Gish W, and Altschul SF (1991) METHODS: A companion to Methods in Enzymology 3:66-70.

Matrix and Gap Costs

- **Matrix**

A key element in evaluating the quality of a pairwise sequence alignment is the "substitution matrix", which assigns a score for aligning any possible pair of residues. The matrix used in a BLAST search can be changed depending on the type of sequences you are searching with (see the [BLAST Frequently Asked Questions](#)). See more information on [BLAST substitution matrices](#).

- **Gap Cost**

The pull down menu shows the Gap Costs for the chosen Matrix. There can only be a limited number of options for these parameters. Increasing the Gap Costs will result in alignments which decrease the number of Gaps introduced.

• PSSM

PSI-BLAST can save the Position Specific Score Matrix constructed through iterations. The PSSM thus constructed can be used in searches against other databases with the same query by copying and pasting the encoded text into the PSSM field.

To save a PSSM file:

- Run a protein BLAST search.
- Check the PSI-BLAST box on formatting page.
- Click the "Format" Button.
- On the PSI-BLAST results page, click the "Run PSI-BLAST Iteration 2" button.
- Now, on the Format page, select "PSSM" from the "Show" pull down menu.
- Click "Format" button.
- This will display text output with the ASCII-encoded PSSM. The "Save as..." option of the browser can be used to save this to a plain text file on your hard drive.

To use the PSSM in a new protein BLAST search against other databases:

- Copy the above PSSM from the browser
- Open a new protein BLAST page
- Paste the PSSM in the PSSM field in the page
- provide the SAME query in the search box
- select a different target database
- click "BLAST" button to start the search

If the database is the same as when the PSSM was stored, you'll reproduce the iteration on which you've saved the PSSM; A different database will yield a different hit list.

NCBI BLAST Advanced Options

Program Advanced Options

Accepted Parameters for Other Advanced Field

- G** Cost to open gap [Integer]: default = 5 for nucleotides/ 11 for proteins
- E** Cost to extend gap [Integer]: default = 2 for nucleotides/ 1 for proteins
- q** Penalty for nucleotide mismatch [Integer]: default = -3
- r** reward for nucleotide match [Integer]: default = 1
- e** expect value [Real]: default = 10
- W** wordsize [Integer]: default = 11 for nucleotides/ 28 for megablast/ 3 for proteins
- y** Dropoff (X) for blast extensions in bits: default = 20 for blastn/ 7 for others
- X** X dropoff value for gapped alignment (in bits): default = 15 for all programs, not applicable to blastn
- Z** final X dropoff value for gapped alignment (in bits): 50 for blastn 25 for others

Only limited values for gap existence and extension are supported for BLAST programs. For protein BLAST, see pulldown menu display next to the Matrix for details. For nucleotide BLAST, see [News on 2.2.13 release](#).

PHI-BLAST Pattern

PHI-BLAST (Pattern-Hit Initiated BLAST) is a search program that combines matching of regular expressions with local alignments surrounding the match. Given a protein sequence *S* and a regular expression pattern *P* occurring in *S*, PHI-BLAST helps answer the question:

What other protein sequences both contain an occurrence of P and are homologous to S in the vicinity of the pattern occurrences?

PHI-BLAST may be preferable to just searching for pattern occurrences because it filters out those cases where the pattern occurrence is probably random and not indicative of homology. See [PHI-BLAST pattern syntax](#) for details.

C. Result Format Options

Graphical Overview

An overview of the database sequences aligned to the query sequence is shown. The score of each alignment is indicated by one of five different colors, which divides the range of scores into five groups. Multiple segments of alignments to the same database sequence are connected by a thin grey line. Mousing over a hit sequence causes the definition and score to be shown in the window at the top, clicking on a hit sequence takes the user to the associated alignments.

NCBI-gi

Checking this option causes NCBI gi identifiers to be shown in the output, in addition to the accession and/or locus name. Examples with and without this are given below.

```
gi|28559089|ref|NM_000249.2| Homo sapiens mutL homolog 1, col... 5003 0.0 UniGene infoGene info
ref|NM_026810.1| Mus musculus mutL homolog 1 (E. col 1344 0.0 UniGene infoGeoGene info
```

format

This determines which object to report and in what format. The default for the object is "Alignment", which can be changed to PSSM, or BioSeq (ANS.1 seqAlign output), with PSSM only available for PSI-BLAST searches. The format of the result appears in the browser is set to "HTML" by default, which can be changed to "plain text", ASN.1, or XML.

CDS feature

Checking this option will allow BLAST formatter to parse out the annotated sequence features found in or around the vicinity of hits and display them within the BLAST result. For custom query sequences, it will also translate the CDS using the CDS translation annotated on matching database sequence as a guide. Mismatch in translation will be highlighted in pink. A representative example with CDS translation is given below.

```
>gi|46452254|gb|AY585334.1| Sus scrofa cystic fibrosis transmembrane conductance regulator
(CFTR) mRNA, complete cds
Length=4449
```

```
Score = 5453 bits (2751), Expect = 0.0
```

Identities = 4036/4449 (90%), Gaps = 6/4449 (0%)
Strand=Plus/Plus

```

CDS: Putative 1      1      M Q R S P L E K A S V V S K L F F S W T
Query               133    ATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGTTGTCTCCAAACnnnnnnnncAGCTGGACC 192
                   |||
Sbjct               1      ATGCAGAGGTCGCCTCTGGAAAAGGCCAGCATCTTCTCCAAACTTTTTTCAGCTGGACC 60
CDS:cystic fibrosis 1      M Q R S P L E K A S I F S K L F F S W T

CDS: Putative 1      21     R P I L R K G Y R Q R L E L S D I Y Q I
Query               193    AGACCAATTTTGAGGAAAGGATACAGACAGCGCCTGGAATTGTCAGACATATACCAAATC 252
                   |||
Sbjct               61     AGACCAATTTTGAGAAAAGGATATAGACAGCGCCTGGAATTGTCAGACATATACCATATC 120
CDS:cystic fibrosis 21     R P I L R K G Y R Q R L E L S D I Y H I

CDS: Putative 1      41     P S V D S A D N L S E K L E R E W D R E
Query               253    CCTTCTGTTGATTCTGCTGACAATCTATCTGAAAAATTGGAAAGAGAATGGGATAGAGAG 312
                   |||
Sbjct               121    TCTTCTTCTGACTCTGCTGACAATCTGTCTGAAAAATTGGAAAGAGAATGGGACAGAGAA 180
CDS:cystic fibrosis 41     S S S D S A D N L S E K L E R E W D R E

```

Masking

There are two options that determines the way filter masked region should be displayed in.

- **Masking Character**

"X or N" displays the masked region in X for protein and N for nucleotide

"Lower Case" displays masked region in lower case letters

- **Masking Color**

The masked region can be "highlighted" with grey or red colored fonts

Descriptions

This option restricts the number of short descriptions of matching sequences reported to the number specified. Default setting varies from page to page. See also EXPECT.

Alignments

This option restricts database sequences to the number specified for which high-scoring segment pairs (HSPs) are reported. Different pages have different default settings. If more database sequences than this happen to satisfy the statistical significance threshold for reporting, only the matches ascribed the greatest statistical significance are reported. See EXPECT below.

Database LinkOuts

Enabling this option provides cross reference links from the BLAST hits to entries found in other specialized databases from NCBI. If a database sequence matches the query and it also included in Gene, UniGene, Structure, or GEO database, you will be able to follow the link to GIF icon marked links to get additional information for that hit.records in those resources.

G = Gene Link **U** = UniGene Link **E** = GEO Link **S** = 3D Link **M** = MapView Link

Sequence Retrieval

Check this box will allow BLAST formatter to display a set of sequence retrieval button, as given in the examples below, that allow user to choose all or a manually selected subset of matched sequences for

downloading through Entrez Nucleotide or Entrez Protein database.

Get selected sequences

Select all

Deselect all

Alignments Views

Standard BLAST alignment in pairs of query sequence and database match. For nucleotide, the matches are marked by a pipe symbol ("|") in between query and database sequence. For protein, the identical matches are marked by letter code with "homologous" substitutions (determined by the scoring matrix used) marked by "+" symbol in a line between the query and the database sequence.

New View This instructs formatter to format the BLAST result using the newly introduced formatting capabilities, which displays the "Description" section in a new tabular format and allow users to re-sort matching database sequences according to the "maximum identity", "percent coverage", "total score", or "max score" of the alignments. For individual HSPs from the same database sequence, users can re-sort them according to their "Score", "Percent identity", "Query start position", and "Subject start position". Matches to entries from the two newly introduced BLAST databases will have links to "MapView" marked by the M icon (M).

- **Pairwise with identity**

The databases alignments are anchored (shown in relation to) to the query sequence in pairwised fashion with mismatches colored in red. "Sbjct" will be in red and bold font if a line in the alignment contains mismatches. See [example](#) below.

- **Query-anchored with identities**

The databases alignments are anchored (shown in relation to) to the query sequence. Identities are displayed as dots (.), with mismatches displayed as single letter abbreviations.

- **Query-anchored without identities**

Identities are shown as single letter nucleotide abbreviations.

- **Flat Query-anchored with identities**

The 'flat' display shows inserts as deletions on the query. Identities are displayed as dots (.), with mismatches displayed as single letter abbreviations.

- **Flat Query-anchored without identities**

The 'flat' display shows inserts as deletions on the query. Identities are shown as single letter abbreviations.

- **Hit Table**

Simple output with different alignment information separated according to tab delimited fields with field headers are displayed at the top.

```
>gi|21536448|ref|NM_002622.3| U E G Homo sapiens prefoldin 1 (PFDN1), mRNA
Length=1296
```

```
Score = 392 bits (212), Expect = 2e-107
Identities = 220/223 (98%), Gaps = 3/223 (1%)
Strand=Plus/Plus
```

Query	107	TCCTACCTGGAGCGAAG-GTTANAGGAAGCTGAGGACAACATCCGGGAGATGCTGATGGC	165
Sbjct	300C.....-.....	358
Query	166	ACGAAGGG-CCAGTAGGGAGCCTCTCTGGGAAGCTCTCCTCCTGCCCTCCCATTCCTG	224
Sbjct	359C.....	418
Query	225	GTGGGGGCAGAGGAGTGTCTGCAGGAAACAGCTTCTCCTCTGCCCGATGGATGCTTTA	284
Sbjct	419	478
Query	285	TTTGGATGGCCTGGCAACATCACATTTTCTGCATCACCCCTGAG	327
Sbjct	479	521

NOTE: Links off the accession and gif icons are disabled.

To get XML, text, or ASN.1 output, click on the HTML next to the "format" and select from the pulldown menu. ASN.1 SeqAnnot format is for importation into NCBI toolkit programs. The imported output can be converted to other display format. See documents under <ftp.ncbi.nlm.nih.gov/blast/demo/> for more information.

Format for PSI-BLAST

The Position-Specific Iterated BLAST (PSI-BLAST) program performs iterative searches with a protein query, in which sequences found in one round of search are used to build a custom score model for the next round.

In PSI-BLAST the algorithm is not tied to a specific score matrix, such as BLOSUM62, which has been implemented using an **AxA** substitution matrix where A is the alphabet size. Instead, it uses a **QxA** matrix, where Q is the length of the query sequence. At each position the cost of a letter depends on the position with regard to the query and the letter in the subject sequence.

To run this search, "Format for PSI-BLAST" checkbox must be checked.

Inclusion Threshold

This sets the statistical significance threshold for including a sequence in the model used by PSI-BLAST to create the PSSM on the next iteration.

Limit results by entrez query

This function is similar to the "Limit by Entrez Query terms" in the option section. The only difference is that it applies only to the identified hits. In another word, it is applied post-search and allows users to see only hits fitting the requirement of the Entrez query terms. Default is to format without input query terms and allow users to see all the hits.

Expect value range

This instructs BLAST formatter to display hits with Expect value within the specified range. Default value is 0 to Expect value setting. Lower bound goes to the first box, higher bound goes to the second box.

layout

This determines whether BLAST report the result in a newly spawned browser window ("two window", as default) or in the same window the initial RID was reported in.

Formatting options on page with results

This instructs BLAST whether to display the format options in the result page or not. Default is not to display such options.

AutoFormat

When the AutoFormat option is selected, clicking the "Submit" button will submit the search and force browser to check for result in a defined schedule and then automatically format BLAST results when they are ready. The default "Semiauto" setting requires user manually hit the "FORMAT" button to initial this process.

If AutoFormat is disabled, clicking the "Format" button in the format page will only make browser check for result once. If the result is not ready, the web page will not be automatically updated. One needs to manually reload the page, or hit "Format" button again.

Results file

Checking this option will instruct browser to save the page content to a file through a dialog box rather than display them in a browser window. This option is only available from MEGABLAST pages.

Get the URL with preset values

This button allows users to save the adjusted the search parameters of a given page. It does so by first captures the changes and put them into a new URL. Users need to follow this URL to get a new page with the adjusted parameters and bookmark the new page to "Save" the page.

D. Rules for pattern syntax for PHI-BLAST

Web PHI-BLAST search requires a pattern along with a protein sequence containing the pattern. A simple example and how to use PHI-BLAST is available from [this page](#).

The syntax for pattern specification in PHI-BLAST follows the conventions of PROSITE. When using the stand-alone program, it is permissible to have multiple patterns in a file separated by a blank line between patterns. When using the Web-page only one pattern is allowed per query.

Accepted PHI-BLAST Pattern Vocabulary	
ABCDEFGHIJKLMNPQRSTVWXYZU	Protein alphabet
ACGT	DNA alphabet
[]	means any one of the characters enclosed in the brackets e.g., [LFYT] means one occurrence of L or F or Y or T
-	nothing, used as a spacer to clearly separate each position
x	with nothing following means any residue
(n)	means the preceeding residue is repeated 5 times
(m, n)	the preceeding residue is repeated between m to n times (n > m)
>	only at the end of a pattern and means nothing it may occur before a period
.	may be used at the end, means nothing

When using the stand-alone program, the pattern should be stored in a pattern input file, with the first line starting with ID followed by 2 spaces and a text string giving the pattern a name. There should also be a

line starting with PA followed by 2 spaces and then the pattern description.

All other PROSITE codes in the first two columns are allowed, but only the HI code, described below is relevant to PHI-BLAST.

Here is an example from PROSITE:

```
ID CNMP_BINDING_2; PATTERN. AC PS00889;
DT OCT-1993 (CREATED); OCT-1993 (DATA UPDATE); NOV-1995 (INFO UPDATE).
DE Cyclic nucleotide-binding domain signature 2.
PA [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].
NR /RELEASE=32,49340;
NR /TOTAL=57(36); /POSITIVE=57(36); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR /FALSE_NEG=1; /PARTIAL=1;
CC /TAXO-RANGE=??EP?; /MAX-REPEAT=2;
```

The line starting with ID gives the pattern a name.

The lines starting with AC, DT, DE, NR, NR, and CC are relevant to PROSITE users, but irrelevant to PHI-BLAST. These lines are tolerated, but ignored by PHI-BLAST.

The line starting with PA describes the pattern, which can be explained as the following.

Pattern Position	Pattern Syntax	Meaning
1	[LIVMF]	one of LIVMF
2	G	G
3	E	E
4	X	any one residue
5	[GAS]	one of GAS
6	[LIVM]	one of LIVM
7	X(5,11)	5 to 11 any residue
8	R	R
9	[STAQ]	one of STAQ
10	A	one A
11	X	any one residue
12	[LIVMA]	one of LIVMA
13	X	any one residue
14	[STACV]	any one of STACV
Note: total length of this motif/pattern is between 18 to 24 residues.		

In this case the pattern ends with a period. It can end with nothing after the last specifying symbol or any number of > signs or periods or combination thereof. Given below is another example, illustrating the use of an HI line.

```
ID ER_TARGET; PATTERN.
PA [KRHQSA]-[DENQ]-E-L>.
HI (19 22)
HI (201 204)
```

In this example, the HI lines specify that the pattern occurs twice, once from positions 19 through 22 in the sequence and once from positions 201 through 204 in the sequence. These specifications are relevant when stand-alone PHI-BLAST is used with the seedp option, in which the interesting occurrences of the pattern in the sequence are specified. In this case the HI lines specify which occurrence(s) of the pattern should be used to find good alignments.

In general, the seedp option is more useful than the standard patternp option ONLY when the pattern

occurs $K > 1$ times in the sequence AND the user is interested in matching to $J < K$ of those occurrences. Then using the HI lines enables the user to specify which occurrences are of interest.

For simple pattern searches, use seedtop from NCBI's standalone command line blast package instead. For more information, see <ftp.ncbi.nlm.nih.gov/blast/documents/seedtop.html>.

This document is also available in [pdf format](#).

[Disclaimer](#)

[Privacy statement](#)

[Accessibility](#)

This page is [valid XHTML 1.0](#).