An official website of the United States government
Here's how you know

Log in

BLAST ®

Home Recent Results Saved Strategies Help

BLAST > blast-help > BLAST Topics

Was this content helpful?

Query Input and database selection

The query sequence(s) to be used for a BLAST search should be pasted in the *'Search'* text area. BLAST 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.

Accepted Input Formats

FASTA

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) 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:

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
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 indeterminate length
```

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

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

Feedback

Halozyme EX2169
Merck v. Halozyme
PGR2025-00050

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

Note:

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

[2] The BLAST webpage will not accept "-" in the query. To represent gaps, use a string of N or X instead.

Bare Sequence

This may be just lines of sequence data, without the FASTA definition line, e.g.:

```
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

It can also be sequence interspersed with numbers and/or spaces, such as the sequence portion of a GenBank/GenPept flatfile report:

```
  1 qikdllvsss tdldttlvlv naiyfkgmwk tafnaedtre mpfhvtkqes kpvqmmcmnn
 61 sfnvatlpae kmkilelpfa sgdlsmlvll pdevsdleri ektinfeklt ewtnpntmek
121 rrvkvylpqm kieekynlts vlmalgmtdl fipsanltgi ssaeslkisq avhgafmels
181 edgiemagst gviedikhsp eseqfradhp flflikhnpt ntivyfgryw sp
```

Blank lines are not allowed in the middle of bare sequence input.

Identifiers

Normally these are simply an accession or accession.version. The identifier may consist of only one token (i.e., word). Spaces between letters in the input will cause it to be treated as bare sequence (spaces before or after the identifier are allowed). Examples of illegal input are:

```
ACCESSION   P01013
AAA68881. 1
gi| 129295
```

For the first example "ACCESSION" must be removed, in the second example there is a space before the version number of the accession, in the third example there is a space after the bar ("|").

If more than one query is specified, each identifier should be on a separate line.

Upload file

This function allows users to upload a text file containing queries formatted in FASTA format. The file can also contain sequence identifiers instead of FASTA sequences.

Query subrange

A segment of the query sequences can be used in BLAST searching. You can enter the range in the "Form" and "To" boxes provided under "Query subrange" to specify the position of this segment. For 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.

Query Genetic Code

Genetic code to be used in blastx and tblastx translation of the query. See list of Genetic Codes in Taxonomy

FOLLOW NCBI

     

Follow NLM

National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894

Copyright
FOIA
HHS Vulnerability Disclosure

Help
Accessibility
Careers

NLM NIH HHS USA.gov