EMBL-EBI

EBI > Help > FASTA

FASTA/SSEARCH/GGSEARCH/GLSEARCH Help

LAUNCH APPLICATION

- INTRODUCTION

FASTA (pronounced FAST-AYE) stands for FAST-All, reflecting the fact that it can be used for a fast protein comparison or a fast nucleotide comparison. This program achieves a high level of sensitivity for similarity searching at high speed. This is achieved by performing optimised searches for local alignments using a substitution matrix. The high speed of this program is achieved by using the observed pattern of word hits to identify potential matches before attempting the more time consuming optimised search. The trade-off between speed and sensitivity is controlled by the ktup parameter, which specifies the size of the word. Increasing the ktup decreases the number of background hits. Not every word hit is investigated but instead initially looks for segment's containing several nearby hits.

FASTA uses four steps to calculate three scores that characterise sequence similarity. These steps are outlined below. A representation of these steps is reported in a postscript format figure drawn from Barton (1994) Protein Sequence Alignment and Database Scanning.

Step 1 : Identify regions shared by the two sequences with the highest density of identities (ktup=1) or pairs of identities (ktup=2).

The first step uses a rapid technique for finding identities shared between two sequences; the method is similar to an earlier technique described by Wilbur and Lipman. FASTA achieves much of its speed and selectivity in this step by using a lookup table to locate all identities or groups of identities between two DNA or amino acid sequences during the first step of the comparison. The ktup parameter determines how many consecutive identities are required in a match. A ktup value of 2 is frequrntly used for protein sequence comparison, which means that the program examines only those portions of the two sequences being compared that have at least two adjacent identical residues in both sequences. More sensitive searches can be done using ktup = 1. For DNA sequence comparisons, the ktup parameter can range from 1 to 6; values between 4 and 6 are recommanded. When the query sequence is a short oliginucleotide of oligopeptude, ktup = 1 should be used.

In conjunction with the lookup table, we use the "diagonal" method to find all regions of similarity between the two sequences, counting ktup matches and penalizing for intervening mismatches. This method identified regions of a diagonal that have the highest densitu of ktup matches. The term diagonal refers to the diagonal line that is seen on a dot matrix plot when a sequence is compared with itself, and it denotes an alignment between two sequenves without gaps. FASTA uses a formula for scoring ktup matches that incorporates the actual PAM250 values for the aligned residues. Thus, groups of identities with high similarity scores contribute more to the local diagonal score than to identities with low similarity scores. This more sensitive formula is used for protein sequence comparisons; the constant value for ktup matches is used for DNA sequence comparisons. FASTA saves the 10 best local regions, regardless of whether they are on the same of different diagonals.

Step 2 : Rescan the 10 regions with the highest density of identities using the PAM250 matrix. Trim the ends of the region to include only those residues contributing to the highest score. Each region is a partial alignment without gaps.

After the 10 best local regions are found in the first step, they are rescored using a scoring matrix that allows runs of identities shorter than ktup residues and conservative replacements to contribute to the similarity score. For protein sequences, this score is usually caculated using the PAM250 matrix, although scoring matrices based on the minimum number of base changes required for a specific replacement, on identities alone, or on an alternative measure of similarity, can also be used with FASTA. The PAM250 scoring matrix was derived from the analysis of the amino acid replacements occuring among related proteins, and it specifies a range of positive scores for replacements that commonly occur among related proteins and negative scores for unlikely replacements. FASTA can also be used for DNA sequence comparisons, and matrices can be constructed that allow separate penalties for transitions and transversions.

For each of the best diagonal regions rescanned with the scoring matrix, a subregion with the maximal score is identified. Initial scores are used to rank the library sequences. These scores are referred to as init1 score.

Step 3 : If there are several initial regions with scores greater than the CUTOFF value, check to see whether the trimmed initial regions can be joined to form an approximate alignment with gaps. Calculate a similarity score that is the sum of the joined initial regions minus a penalty (usually 20) for each gap. This initial similarity score (initn) is used to rank the library sequences. The score of the single best initial region found in step 2 is reported (init1).

FASTA checks, during a library search, to see whether several initial regions can be joined together in a single alignment to increase the initial score. FASTA calculates an optimal alignment of initial regions as a combination of compatible regions with maximal score. This optimal alignment of initial regions can be rapidily calculated using a dynamic programming algorithm. FASTA uses the resulting score, referred to as the initn score, to rank the library sequences. The third "joining" step in the computation of the initial score increases the sensitivity of the search method because it allows for insertions and deletions as well as conservative replacements. The modification does, however, decrease selectivity. The degradation selectivity is limited by including in the optimization step only those initial regions whose scores are above an empirically determined threshold : FASTA joins an initial region only if its similarity score is greater than the cutoff value, a value that is approximately one standard deviation above the average score expected from unrelated sequences in the library. For a 200-residue query sequence and ktup-2, this value is 28.

Step 4 : constructs NWS (Needleman-Wunch-Sellers algorithm) optimal alignment of the query sequence and the library sequence, considering only those residues that lie in a band 32 residues wide centered on the best initial region found in Step 2. FASTA reports this score as the optimized (opt) score. After a complete search of the library, FASTA plots the initial scores of each library sequence in a histogram, calculates the mean similarity score for the query sequence against each sequence in the library, and determines the standard deviation of the distribution of initial scores. The initial scores are used to rank the library sequences, and, in the fourth and final step of the comparison, the highest scoring library sequences are aligned using a modification of the standard NWS optimization method. The optimization employs the same scoring matrix used in determining the initial regions; the resulting optimized alignments are calculated for further analysis of potential relationships, and the optimized similarity score is reported.

Lookup table

A lookup table is a rapid method for finding the position of a residue in a sequence. One way to find the "A" in the sequence "NDAPL" is to compare "A" to each residue in the sequence. A faster way, is to make a table of all possible residues (23 for proteins) so that the computer representation for the residue (i.e "A" is 1, "R" is 2, "N" is 3) is the same as its position in the table. A value is then placed in the table that indicates whether the residue is present in the sequence and, if it is, where it is present. For this example the table has the value 1 at position 3, 2 at position 4, 3 at position 1, 4 at 15, 5 at 11, and the remainning 18 positions are 0. The position of the "A" in the sequence can then be determined in a single step by looking it up at position 1 in the table.

- **Tutorials**

  2can Nucleotide Similarity Search

  2can Protein Similarity Search
- **Other sources of information**
  1. http://fasta.bioch.virginia.edu/
  2. Original Documentation
  3. fasta3x.txt
  4. More information on the fasta algorithm

- FASTA SEQUENCE FORMAT

  - This format contains a one line header followed by lines of sequence data.
  - Sequences in fasta formatted files are preceded by a line starting with a" >" symbol.
  - The first word on this line is the name of the sequence. The rest of the line is a description of the sequence.
  - The remaining lines contain the sequence itself.
  - Blank lines in a FASTA file are ignored, and so are spaces or other gap symbols (dashes, underscores, periods) in a sequence.
  - FASTA files containing multiple sequences are just the same, with one sequence listed right after another. This format is accepted for many multiple sequence alignment programs.

```
>FOSB_MOUSE Protein fosB. 338 bp
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL
```

- YOUR SEQUENCES

You can cut and paste or type a sequence into the large text window. A free text (raw) sequence is simply a block of characters representing a DNA/RNA or Protein sequence. You may also paste a sequence in GCG, FASTA, EMBL, GenBank, PIR, NBRF, Phylip or UniProtKB/Swiss-Prot format. Partially formatted sequences will not be accepted. Copying and Pasting directly from word processors may yield unpredictable results as hidden/control characters may be present. Adding a return to the end of the sequence may help certain applications understand the input. Some examples of common sequence formats may be seen here.

IF your sequence is DNA and contains more than 50% ambiguity codes it will be rejected by the system (both email and interactive submissions). It is best to use the part(s) of the sequence that contain real DNA (good consensus regions) as the likelihood if these regions giving positive scores is higher than if they contained many 'N's which the programs will attempt to match to the entire database(s). If successful, these database scans will contain false positive hits which are of limited use.

SHORT SEQUENCES

For very short nucleic acid sequences the strategy recommended is to decrease the word length (ktup) from 6 to 1. In this way a significant increase in sensitivity is achieved without loss of biological significance. FASTA may not return as many hits as blast does in this type of searches but the relevance of fasta results is much higher than with blast.
For Fastf3 and Fasts3 you must use the following format to enter the fragments:

```
>mgstm1
MGCEN,
MIDYP,
MLLAY,
MLLGY
```

- YOUR EMAIL

You must type your email address in this text box, it must be a valid internet email address in the form joe@bio.med.org. It is not necessary to fill in the box if you are running your search interactively, where your results will be delivered to the browser window when they are ready.

- SEARCH TITLE

  You may type any text you want to help you identify your search results.

- RESULTS

  This option lets you choose between email and interactive runs. The email run requires you to type an email address in the email text box, For example: joe@somewhere.domain.country, as a link to your results will be delivered by email. The default value is email. You will be delivered your results to your browser, when they become available with an interactive job.

- PROGRAM

  PLEASE READ CAREFULLY!

  The programs available and their uses:

Program	Function	Submission Type
FASTA	scan a protein or DNA sequence library for similar sequences	interactive/email
FASTX/Y	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.	interactive/email
TFASTX/Y	compares a protein to a translated DNA data bank	interactive/email
FASTS	compares linked peptides to a protein databank	interactive/email
FASTF	compares mixed peptides to a protein databank	interactive/email
SSEARCH	scan a protein or DNA sequence library for similar sequences	interactive/email
GGSEARCH	compares a protein or DNA sequence to a sequence database producing global-global alignment (Needleman-Wunsch).	interactive/email
GLSEARCH	compares a protein or DNA sequence to a sequence database	interactive/email

- DATABASES

  Choose here the databases you which to run your protein sequence against. You can choose multiple databases by clicking on them. The choices will appear highlighted. Please note that Netscape and Internet Explorer behave differently when doing this. To choose multiple databases in Netscape simply click on the database names. To deselect a database simply click on it again.
  With Explorer you have to press SHIFT and click on the databases in order to make a multiple choice.
  The database names refer to the following:

Abbreviation	Database Name
	Protein FASTA

UniProt Knowledgebase	The UniProt Knowledgebase (UniProtKB) is the central access point for extensive curated protein information, including function, classification, and cross-references. Search UniProtKB to retrieve "everything that is known" about a particular sequence.
UniProt Clusters	The UniProt Reference Clusters (UniRef) databases combine closely related sequences into a single record to speed searches. There are three different non-redundant databases with different sequence identity cut-offs. In UniRef100, UniRef90 and UniRef50 databases no pair of sequences in the representative set has >100%, >90% or >50% mutual sequence identity. The three UniRef databases allow the user to choose between a fast search and a truly comprehensive one.
UniProt Archive	The UniProt Archive (UniParc) contains available protein sequences collected from many different sources. The sequence data are archived to facilitate examination of changes to sequence data. Search UniParc if you want to examine the "history" of a particular sequence.
UniProtKB/Swiss-Prot	UniProtKB/Swiss-Prot is the manually curated subsection of the UniProt Knowledgebase.
International Protein Index	The International Protein Index (IPI) provides non-redundant proteome sets for a selection of higher eukaryotes, e.g. Arabidopsis, Chicken, Mouse, Human, etc. Cross-references are provided to the various source databases.
Protein Structure Sequences	Protein sequences from structures described in the Brookhaven Protein Data Bank (PDB)
Structural Genomics Targets	Structural Genomic Targets (SGT) database
IntAct *New*	The IntAct sequence database is derived from UniProt entries and data from MassSpec experiments submitted to the IntAct protein-interaction database.
IMGT/HLA	The human major histocompatibility complex (HLA) section of the the international immunogenetics (IMGT) database.
EPO Patent Protein Sequences	Protein sequences appearing in patents from the European Patent Office (EPO)
JPO Patent Protein Sequences	Protein sequences appearing in patents from the Japanese Patent Office (JPO)
KIPO Patent Protein Sequences	Protein sequences appearing in patents from the Korean Intellectual Property Office (KIPO)
USPTO Patent Protein Sequences	Protein sequences appearing in patents from the United States Patent and Trademark Office (USPTO)

UniProt Clusters 100% (SEG filtered)	UniProt Reference Clusters database (SEG filtered) with entries that have 100% mutual sequence identity.
Nucleotide FASTA	
EMBL Release	The quarterly release of the EMBL nucleotide sequence databank (EMBL-Bank).
EMBL Updates	Daily updates to the quarterly EMBL nucleotide sequence databank (EMBL-Bank) release.
EMBL Coding Sequence	The nucleotide sequences of the coding sequence (CDS) features in the EMBL nucleotide sequence databank (EMBL-Bank).
EMBL *subsets*	Sequences from the EMBL nucleotide sequence databank (EMBL-Bank) classified by data class and/or taxonomic division. For example: EMBL EST Environmental contains sequences where data class is EST and taxonomic division is Environmental.
EMBL Vectors	Sequencing vectors extracted from the EMBL nucleotide sequence databank (EMBL-Bank)
IMGT/LIGM-DB	The immunoglobulins and T cell receptors (LIGM-DB) section of the international immunogenetics (IMGT) database.
IMGT/HLA	The human major histocompatibility complex (HLA) section of the the international immunogenetics (IMGT) database.
HGVBASE	A database of human sequence variations.
HGBASE	European SNP database
Alternative Splicing Database (ASD) FASTA	
AEDB Exons	Sequences of manually collected alternative exons.
AltSplice Genes	Sequences of genes for which AltSplice has confirmed alternative events.
AltSplice Isoforms	Sequences of splice patterns for genes of which AltSplice has confirmed alternative events.
ASD Peptides	ASD Peptides database.
Ligand Gated Ion Channel Databases FASTA	
LGICdb Protein	Protein sequence similarity searching against the Ligand Gated Ion Channel Database.
LGICdb Nucleotide	Nucleotide sequence similarity searching against the Ligand Gated Ion Channel Database.

- MATRIX

  Use this option to set which comparison matrix should be used when searching the database. The default matrix for blast is blosum62. You may choose from a complete list of matrices which should cover various evolutionary constraints. More on matrices.

- GAP PENALTIES

  GAPOPEN: Penalty for the first residue in a gap (-12 by default for fasta with proteins, -16 for DNA).
  GAPEXT: Penalty for additional residues in a gap (-2 by default for fasta with proteins, -4 for DNA).

  More on Gaps.

- SCORES

  Setting this option to any number available in the menu allows you to set to maximum number of reported scores in the output file.

  Example of a score:

```
 UNIPROT:FOSB_MOUSE P13346 Protein fosB.              ( 338) 2268  354 3.3e-96
```

Database	UniProt I.D.	InterPro I.D.	Description	Length (B.P.)	opt	bits	E-Value
UNIPROT	FOSB_MOUSE	P13346	Protein fosB.	(338)	2268	354	3.3e-96

  For DNA/RNA [f]stands for forward and [r]for reverse, and represents which strandfrom which the alignment was based on.

- ALIGNMENTS

  Setting this options to any number available in the menu allows you to set the maximum number of reported alignments in the output file.

  Note that matching sequences are connected with a " |" symbol. As this is a perfect match, all of the nucleotides are connected with a "|" symbol, mismatches would be connected with a space. A gap would be represented with a " -" symbol. Thus a sequence alignment can be represented in the format...

```
AATCCTTGAGCA
 |    ||||
TAG--ATGAGTT
```

  Example of an alignment:

```
>>UNIPROT:FOSB_MOUSE P13346 Protein fosB.                  (338 aa)
 initn: 2268 init1: 2268 opt: 2268  Z-score: 1866.3  bits: 353.7 E():
3.3e-96
Smith-Waterman score: 2268;  100.000% identity (100.000% ungapped) in
338 aa overlap (1-338:1-338)

               10        20        30        40        50        60
FOSB_M MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
               10        20        30        40        50        60

               70        80        90       100       110       120
```

```
FOSB_M ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
               70        80        90       100       110       120

              130       140       150       160       170       180
FOSB_M GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
              130       140       150       160       170       180

              190       200       210       220       230       240
FOSB_M DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
              190       200       210       220       230       240

              250       260       270       280       290       300
FOSB_M LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
UNIPRO LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
              250       260       270       280       290       300

              310       320       330
FOSB_M TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL
       ::::::::::::::::::::::::::::::::::::::
UNIPRO TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL
              310       320       330
```

- KTUP

  Change this value to limit the word-length the the search should use. A word-length of 2 is sensitive enough for most protein database searches. The thumb rule is that the larger the word-length the less sensitive, but faster the search will be. For DNA searches a ktup of 6 is the default. **Please note that if you do not specify a ktup larger than 3 when doing a nucleic database search ktup will be set automatically to 6.**

- STRAND

  This option lets you choose which DNA strand to search with when you are using a DNA sequence to compare against the DNA databanks. The 'default' is to search the 'both' strands. 'top' means the sequence will be searched as it is input into the form. 'bottom' means: reverse and complement your input sequence.
  More about strands.

- **HISTOGRAM**

  Setting this option to "yes" will display the search histogram of the expected frequency of chance occurrence of the database matches found. It provides you with a way of quickly checking to see if your statistical estimates are as you might expect. The histogram presents observed and expected distribution of E values.

  Example of a histogram:

```
       opt      E()
< 20  1040     0:=
  22     0     0:           one = represents 1534 library sequences
  24     4     1:*
  26    14    19:*
  28    53   201:*
  30   227  1223:*
  32  1161  4728:=  *
  34  5065 12823:====    *
  36 16019 26336:===========      *
  38 33416 43523:======================      *
  40 58656 60711:=======================================*
  42 81562 74211:================================================*=====
  44 87125 81862:=====================================================*===
  46 92000 83378:======================================================*=====
  48 83023 79825:====================================================*==
  50 83389 72841:===============================================*=======
  52 68683 64039:=========================================*===
  54 58489 54701:===================================*===
  56 48841 45692:=============================*==
  58 36330 37512:========================*
  60 26755 30387:================== *
  62 21306 24361:============== *
  64 17453 19374:============*
  66 13701 15313:=========*
  68 10436 12045:=======*
  70  8222  9439:======*
  72  6047  7376:====*
  74  5090  5751:===*
  76  4048  4476:==*
  78  3338  3479:==*
  80  2761  2701:=*
  82  2240  2067:=*
  84  1752  1637:=*
  86  1600  1267:*=
  88  1288   980:*          inset = represents 22 library sequences
  90  1027   758:*
  92  1066   587:*         :==========================*=============
  94   731   454:*         :====================*=============
  96   707   351:*         :===============*=================
  98   670   272:*         :============*==================
 100   611   210:*         :=========*==================
 102   627   163:*         :=======*=====================
 104   369   126:*         :=====*===========
 106   224    97:*         :====*======
 108   193    75:*         :===*=====
 110   136    58:*         :==*====
 112    86    45:*         :==*=
 114    67    35:*         :=*==
 116    52    27:*         :=*=
 118    41    21:*         :*=
>120   360    16:*         :*================
```

- EXPECTATION VALUE UPPER LIMIT

  Here you may set the expectation value upper limit for score and alignment display. Generally, in evaluating the E() scores, the following rules of thumb can be used, sequences with E() less than 0.01 are almost always found to be homologous, sequences with E() between 1 and 10 frequently turn out to be related as well.

The defaults are 10.0 for FASTA with protein searches, 5.0 for translated DNA/protein comparisons, and 2.0 for DNA/DNA searches.

- EXPECTATION VALUE LOWER LIMIT

  Expectation value lower limit for score and alignment display. A value of 1e-6 prevents library sequences with E()- values lower than 1e-6 from being displayed. This allows the use to focus on more distant relationships. Thus with this option if set will filter out the best matches and allow more distant relationships to be displayed.

  The default setting for this is zero.

- SEQUENCE RANGE

  This options allows the user to denote which region within the query sequence should be searched. For example, if you submit a sequence of 380 amino acids/nucleotides, you may wish to search the database using the region comprising positions 50 to 200. In this case, the user should type into the text field the numbers: **50-200**.

  The default is to search using the whole query sequence.

- DATABASE SEQUENCE RANGE TO SEARCH

  This option is similar to the above except that it sets the sequence range to search within the database. If the user wishes to search all entries with no more than 300 aa/nt he/she must type **1-300** in the text window. It is also possible to use ranges such as 1000-3000 which indicates sequences with at least 1000 aa/nt and up-to, but no more than 3000 aa/nt.

  The default is to search against the whole database entry.

- MOLECULE TYPE

  This option is used to choose or enforce the molecule type of the query in use for a search. This is useful when using programs such as tfastx and tfasty. Please note that this option need not to be changed when using the standard fasta3 program. More about types of molecules.

- FILTER

  Filtering can eliminate statistically significant but biologically uninteresting reports from the blast output by masking out various segments of the query sequence for regions which are non-specific for sequence similarity searches. This leaves the more biologically interesting regions of the query sequence available for specific matching against database sequences. For example, it may be desired to mask acidic, basic or proline-rich segments of a protein that would otherwise yield overwhelming amounts of uninteresting, non-specific matches against a wide array of protein families. The **SEG** program (Wootton and Federhen, 1993) masks low compositional complexity regions, while **XNU** (Claverie and States, 1993) masks regions containing short-periodicity internal repeats. **SEG+XNU** will combine the above two. The **DUST** program by Tatusov and Lipman can only be used with DNA searches and will mask simple repeats in DNA/RNA sequences.

  The default not to use a filter.

  N.B. "If you have UniProt Clusters 100% (SEG filtered)" selected , you will not be able to set a filter as a filter is already applied.

- STATISTICAL ESTIMATES

  None turns off statistical calculations.
  Regress (the default) uses a weighted regression of average score vs library sequence length;
  MLE uses maximum likelihood estimates of Lambda and K Altshul-Gish uses Altschul-Gish parameters (Altschul and Gish, 1996)

Regress/MLE shuf. estimate the statistical parameters from shuffled copies of each library sequence.

statistical estimates

-z -1,0,1,2,3,4,5

> -z -1 turns off statistical calculations. z 0 estimates the significance of the match from the mean and standard deviation of the library scores, without correcting for library sequence length. -z 1 (the default) uses a weighted regression of average score vs library sequence length; -z 2 uses maximum likelihood estimates of Lambda and K; -z 3 uses Altschul-Gish parameters (Altschul and Gish, 1996); -z 4 -5 uses two variations on the -z 1 strategy. -z 1 and -z 2 are the best methods, in general.

-z 11,12,14,15

> estimate the statistical parameters from shuffled copies of each library sequence. This doubles the time required for a search, but allows accurate statistics to be estimated for libraries comprised of a single protein family.

- SEQUENCE INPUT WINDOW

  You can cut and paste or type a sequence into the large text window. A free text (raw) sequence is simply a block of characters representing a DNA/RNA or Protein sequence. You may also paste a sequence in GCG, FASTA, EMBL, GenBank, PIR, NBRF, Phylip or UniProtKB/Swiss-Prot format. Partially formatted sequences will not be accepted. Copying and Pasting directly from word processors may yield unpredictable results as hidden/control characters may be present. Adding a return to the end of the sequence may help certain applications understand the input. Some examples of common sequence formats may be seen here.

- UPLOAD A FILE

  You may upload a file from your computer which containing a valid sequence in any format (GCG, FASTA, EMBL, GenBank, PIR, NBRF, Phylip or UniProtKB/Swiss-Prot) using this option. Please note that this option only works with Netscape Browsers or Internet Explorer version 5 or later. Some word processors may yield unpredictable results as hidden/control characters may be present in the files. It is best to save files with the Unix format option to avoid hidden windows characters. Some examples of common sequence formats may be seen here.
- OTHER SERVICES:

  This services is also available as an application from the EBI's srs server: http://srs.ebi.ac.uk/

- REFERENCES

Mackey A.J., Haystead T.A., Pearson W.R. (2002)
Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences.
Molecular and Cellular Proteomics 1(2): 139-147.
abstract

Pearson W.R. and Lipman D.J. (1988)
Improved Tools for Biological Sequence Comparison.
PNAS 85:2444- 2448.
abstract

Pearson W.R. (1990)
Rapid and Sensitive Sequence Comparison with FASTP and FASTA.

Methods in Enzymology 183: 63-98.
abstract

Pearson W.R. and Lipman D.J. (1988)
Improved Tools for Biological Sequence Comparison.
Proceedings of the National Academy of Sciences 85: 2444-2448.
abstract

Pearson W.R. (1990)
Rapid and Sensitive Sequence Comparison with FASTP and FASTA.
Methods in Enzymology 183: 63-98.
abstract

Contact:

For Support on this service: Please contact EBI support at http://www.ebi.ac.uk/support/

The Author:
William R. Pearson (email: wrp@virginia.edu)
Department of Biochemistry
Box 440, Jordan Hall
U. of Virginia
Charlottesville, VA

- EXAMPLE

Nucleotide tutorial
Protein tutorial