Questions

1. Using the Needleman and Wunsch dynamic programming method, construct the partial alignment score table for the following two sequences, using the scoring parameters: match score: +1; mismatch score: 0 and gap penalty: -1

   ACAGTCGAACG and ACCGTCCG

2. Using the Smith-Waterman method, construct the partial alignment scoring table for a local alignment of the following two sequences:

   ACGTATCGCGTATA and GATGCTCTCGGAAA

   scoring parameters: match score: +1; mismatch score: 0 and gap penalty: -1

Database searches

The common use for sequence alignments is to search through a database of many sequences to retrieve similar sequences to the query sequence.

E.g. we have a region of human genome with unidentified function

Search with millions of other sequences in Genbank of NCBI (or EBI or DDBJ)

Get the regions that align well with the query sequence

Compare with their functional role

Information on regulation/expression and its relationship with other genes.

Major points:

1. Size of the query sequence

2. Number of sequences in the database

Proper methods are necessary to identify correct sequence matches.

BLAST

BLAST: Basic Local Alignment Search Tool

BLAST finds sub-sequences from a sequence database for any query sequence.

Program name	Query sequence	Database type
Blastp	Protein	Protein
Blastn	Nucleic acid	Nucleic acid
Tblastn	Protein	Nucleic acid (translated)
Tblastx	Nucleic acid (translated)	Nucleic acid (translated)

Blastp: searches for protein sequence matches using PAM or BLOSSUM matrices to score the ungapped alignments.

BLAST: Example

Blastp first breaks down the query sequence into words or subsequences of fixed length

All possible pairs are calculated using sliding windows

E.g. AILVPTVI -> AILV, ILVP, LVPT, VPTV and PTVI

Search for word matches (also called High Scoring Pairs or HSPs):

MVQGTIPKLHAILVGTVIAML …

AILVPTVI

Extend the match until the local alignment score falls below a fixed threshold

It also allows gaps in the extended length.

FASTA

FASTA is another program for sequence similarity search and sequence alignment.

FASTA breaks the words into 4-6 nucleotides or 1-2 amino acids

Eg. Query sequence: FAMLGFIKYLPGCM

Word	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Position	2	13			1	5		7	8	4	3		11							9
					6	12				10	14									

Target sequence: TGFIKYLPGACT

1	2	3	4	5	6	7	8	9	10	11	12
T	G	F	I	K	Y	L	P	G	A	C	T
	3	-2	3	3	3	-3	3	-4	-8	2	
	10	3				3		3			

Large number of sequences have the number 3;

Offsetting with 3 gives a reasonable alignment

```
FAMLGFIKYLPGCM
 |||||||
TGFIKYLPGACT
```

Alignment score and statistical significance

The primary indicator of how similar the search results are to a query sequence is the alignment score (S).

Score is given with P or E value.

E-value is the expected number of sequences of score ≥ S that would be found by random choice

P-value is the probability that one or more sequences of score ≥ S would have been found randomly.

Low values of E and P indicate that the search result was unlikely to have been obtained by random chance, and thus is likely to bear an evolutionary relationship to the query sequence.

E values of less than 10^{-3} are often considered indicative of statistically significant results and search algorithms produce matches with E values on the order of 10^{-50}.

BLAST: Features

(i) identifying protein sequences similar to the query,

(ii) finding members of a protein family or build a custom position-specific scoring matrix,

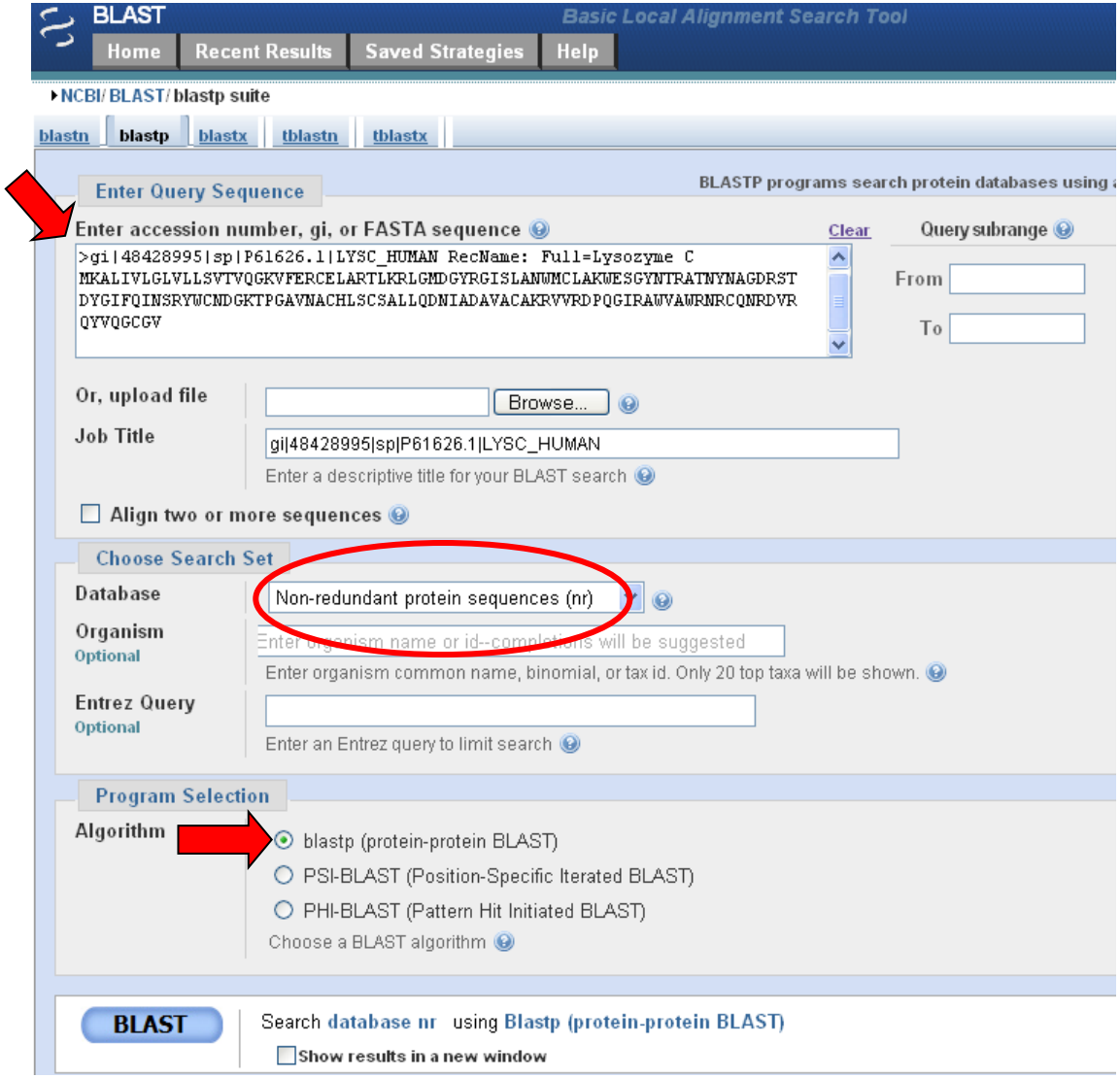
(iii) finding proteins similar to the query around a given pattern,

(iv) finding conserved domains in the query and

(v) searching for peptide motifs.

BLAST is available at http://www.ncbi.nlm.nih.gov/BLAST/.

BLAST: Search



BLAST: Options

Accepts gi number or FASTA format

gi : bar separated NCBI sequence identifier (e.g., gi|48428995).

Accession number : number allotted in Uniprot for each sequence (e.g. P61626)

FASTA format

begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol at the beginning.

Example sequence in FASTA format:

> LYSC_HUMAN RecName: Full=Lysozyme
CMKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESG
YNTRATNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVAC
AKRVVRDPQGIRAWVAWRNRCQNRDVRQYVQGCGV

File formats

1. FASTA format

```
> LYSC_HUMAN RecName: Full=Lysozyme
    CMKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRSTDYG
    IFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRDVRQYVQGCG
    V
```

Files in FASTA format have the extension ".fasta"

2. NBRF/PIR (National Biomedical Research Foundation/Protein Information Resource)

First line begins with >P1 for protein sequence or >N1 for nucleic acid sequence.

```
>P1; LYSC_HUMAN
    CMKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRSTDYG
    IFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRDVRQYVQGCG
    V

*
```

Files in NBRF/PIR format have the extension ".seq" or ".pir"

3. GDE format, (essentially the same as FASTA, the difference is starts with %). The file format is ".gde

All three file formats ignore spaces and carriage returns.

Searchable databases

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.

Uniprot

Last major release of the Uniprot protein sequence database.

pat

Proteins from the Patent division of GenPept.

pdb

Sequences derived from the 3-dimensional structure from Protein Data Bank

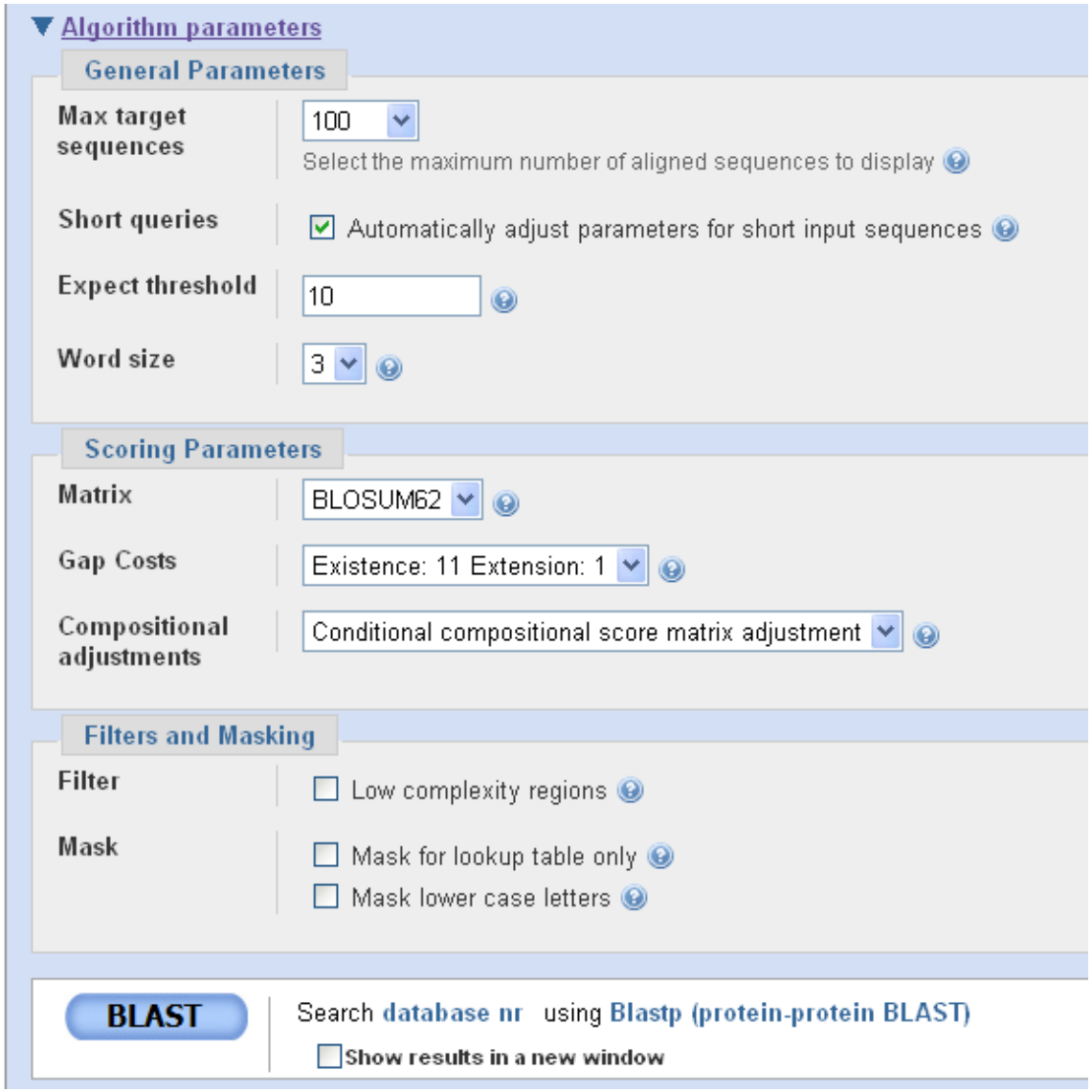
month

All new or revised GenBank CDS translation+PDB+SwissProt+PIR+PRF released in the last 30 days.

env_nr

Metagenomic Protein sequences.

Algorithm parameters



Algorithm parameters

Displaying maximum number of aligned sequences

Expect threshold (e-value): expected number of chance matches in a random model and it is set to 10 as default value.

Word size: length of the seed that initiates an alignment.

scoring parameters can be selected for matrix, gap cost and compositional adjustments.

Substitution matrix: It is a key element in evaluating the quality of a pairwise sequence alignment, which assigns a score for aligning any possible pair of residues.

Generally BLOSUM62 is used as the substitution matrix, which is a 20x20 matrix obtained for all possible substitutions of 20 amino acid residues. It is based on a likelihood method by estimating the occurrence of each possible pairwise substitution using the biochemical character of amino acid residues (aliphatic, aromatic, positive charged, negative charged; polar, sulfur containing).

The gap cost is a cost to create and extend a gap in an alignment.

Further, options are available to filter the low complexity regions and mask query and lower case letters in the sequence.

BLAST: Output

gi|48428995|sp|P61626.1|LYSC_HUMAN RecName:...

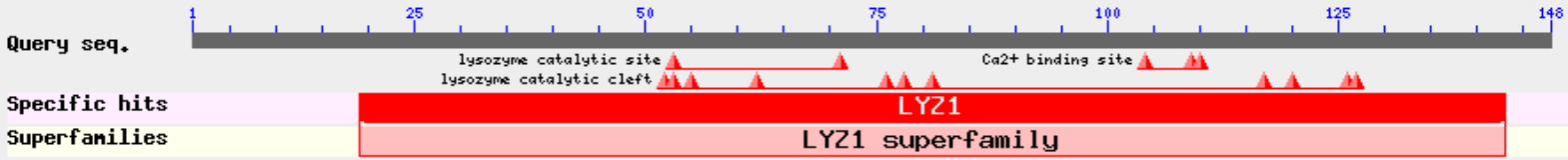
Query ID	lcl|55010	**Database Name**	nr
Description	gi|48428995|sp|P61626.1|LYSC_HUMAN RecName: Full=Lysozyme C	**Description**	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid		
Query Length	148	**Program**	BLASTP 2.2.25+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment]

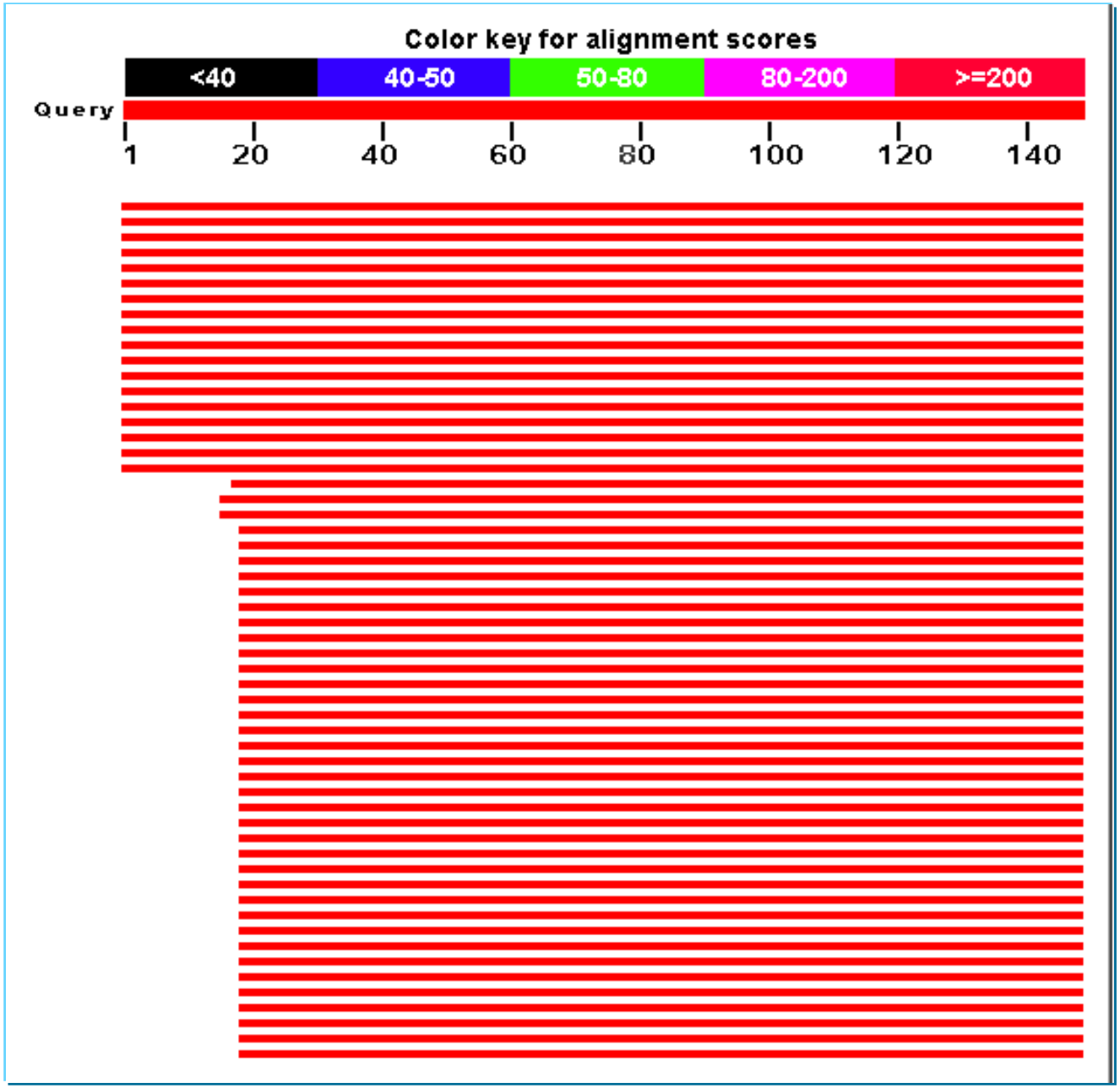
▼ Graphic Summary

▼ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Lysozyme catalytic site
Ca^{2+} binding site


Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
20
40
60
80
100
120
140

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
NP_000230.1	lysozyme C precursor [Homo sapiens] >ref|NP_001009073.1| ly	303	303	100%	4e-81	U G M
AAA36188.1	lysozyme precursor (EC 3.2.1.17) [Homo sapiens]	303	303	100%	6e-81	G M
BAG73364.1	lysozyme [synthetic construct]	301	301	100%	2e-80	
P79179.1	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm	301	301	100%	2e-80	
XP_002823550.1	PREDICTED: lysozyme C-like [Pongo abelii] >sp|P79239.1|LYSC	300	300	100%	5e-80	G M
XP_003259554.1	PREDICTED: lysozyme C-like [Nomascus leucogenys]	298	298	100%	1e-79	G M
AAC63078.1	lysozyme precursor [Homo sapiens]	297	297	100%	4e-79	G M
P79180.1	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm	295	295	100%	2e-78	
P61633.1	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm	283	283	100%	5e-75	
NP_001095203.1	lysozyme C precursor [Macaca mulatta] >sp|P30201.1|LYSC_M	280	280	100%	6e-74	U G M
P79811.1	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm	280	280	100%	6e-74	
NP_001106112.1	lysozyme C precursor [Papio anubis] >sp|P61629.1|LYSC_PAPA	279	279	100%	8e-74	U G
P79806.1	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm	279	279	100%	8e-74	
P79847.1	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm	278	278	100%	1e-73	
P67979.1	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm	278	278	100%	3e-73	
P67977.1	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm	278	278	100%	2e-73	
P79687.1	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm	277	277	100%	4e-73	
P61631.1	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylm	276	276	100%	5e-73	
1C46_A	Chain A, Mutant Human Lysozyme With Foreign N-Terminal Res	272	272	88%	1e-71	S
1C7P_A	Chain A, Crystal Structure Of Mutant Human Lysozyme With Fc	272	272	89%	1e-71	S
CAA53144.1	lysozyme [synthetic construct] >gb|AAQ72808.1| lysozyme [s	271	271	87%	3e-71	
1IOC_A	Chain A, Crystal Structure Of Mutant Human Lysozyme, Eaea-I	271	271	89%	3e-71	S
1LZS_A	Chain A, Structural Changes Of The Active Site Cleft And Diffe	270	270	87%	4e-71	S
1GBW_A	Chain A, Crystal Structure Of Mutant Human Lysozyme Substit	270	270	87%	5e-71	S
1GB6_A	Chain A, Crystal Structure Of Mutant Human Lysozyme Substit	270	270	87%	5e-71	S

Alignment score

Maxscore: Bit score of high scoring pairs (HSPs), similar to Expect Value

Total score: sum of the scores of all HSPs from the same database sequence.

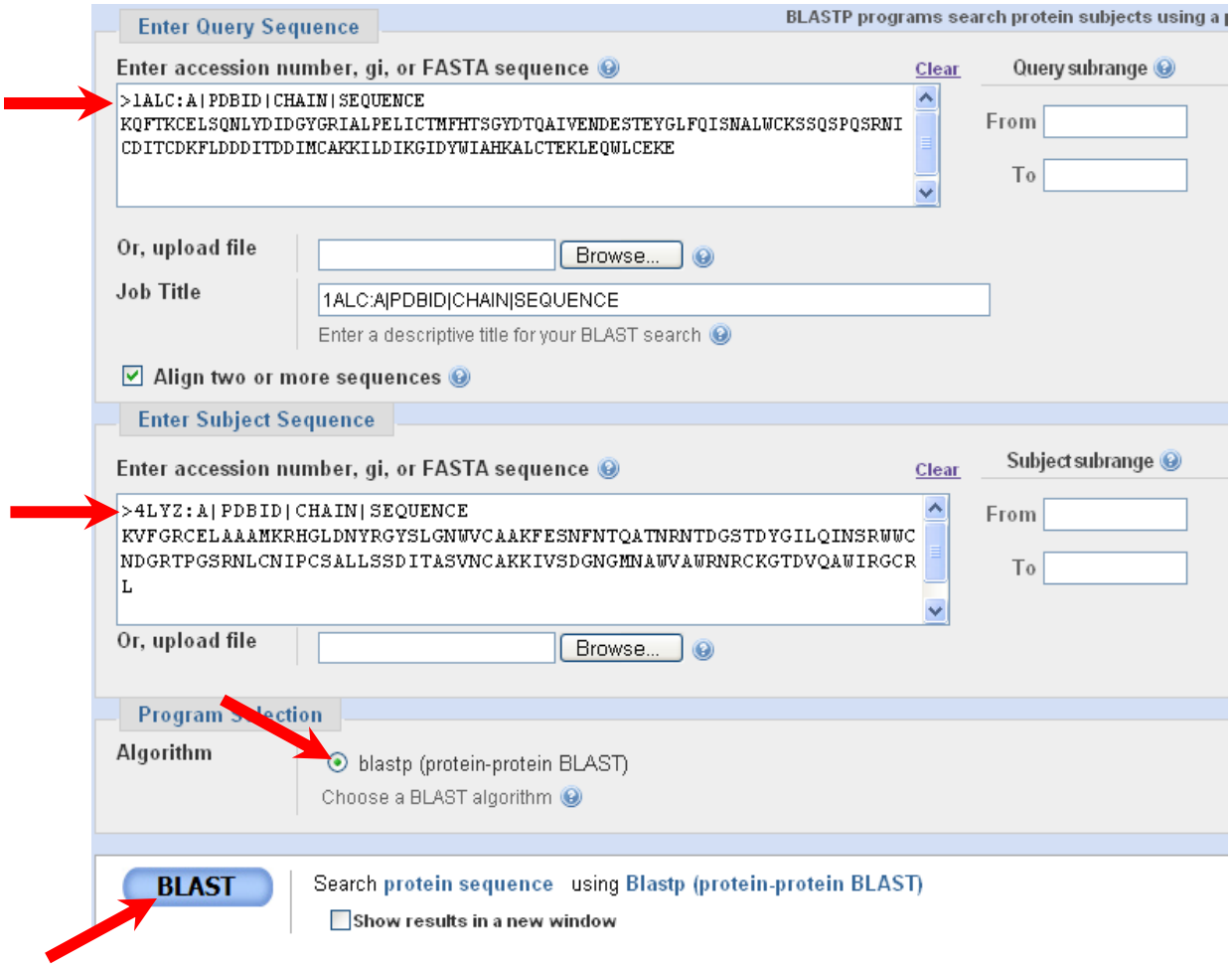
Query Coverage: By the percent of length coverge for the query

E-value: expected number of chance matches in a random model

Alignment scores

```
>gb|AAA36188.1| lysozyme precursor (EC 3.2.1.17)
Length=148

 GENE ID: 4069 LYZ | lysozyme (renal amyloidosis) [Homo sapiens]
(Over 10 PubMed links)

 Score =  303 bits (775),  Expect = 4e-81, Method: Compositional matrix adjust.
 Identities = 147/148 (99%), Positives = 148/148 (100%), Gaps = 0/148 (0%)

Query  1    MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRA  60
            MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRG+SLANWMCLAKWESGYNTRA
Sbjct  1    MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGMSLANWMCLAKWESGYNTRA  60

Query  61   TNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRD  120
            TNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRD
Sbjct  61   TNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRD  120

Query  121  PQGIRAWVAWRNRCQNRDVRQYVQGCGV  148
            PQGIRAWVAWRNRCQNRDVRQYVQGCGV
Sbjct  121  PQGIRAWVAWRNRCQNRDVRQYVQGCGV  148
```

Sequence similarity



Sequence similarity

Blast 2 sequences

1ALC:A|PDBID|CHAIN|SEQUENCE

Query ID	lcl|62163	Subject ID	62165
Description	1ALC:A|PDBID|CHAIN|SEQUENCE	Description	4LYZ:A|PDBID|CHAIN|SEQUENCE
Molecule type	amino acid	Molecule type	amino acid
Query Length	123	Subject Length	129
		Program	BLASTP 2.2.21+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports]

▸ **Graphic Summary**

▸ **Dot Matrix View**

▿ **Descriptions**

```
                                                          Score     E
Sequences producing significant alignments:              (Bits)  Value

lcl|62165  4LYZ:A|PDBID|CHAIN|SEQUENCE                    89.4    2e-23
```

▿ **Alignments** Select All Get selected sequences NEW

```
>lcl|62165 4LYZ:A|PDBID|CHAIN|SEQUENCE
Length=129

 Score = 89.4 bits (220),  Expect = 2e-23, Method: Compositional matrix adjust.
 Identities = 44/115 (38%), Positives = 67/115 (58%), Gaps = 4/115 (3%)

Query  1    KQFTKCELSQNL--YDIDGYGRIALPELICTMFHTSGYDTQAIVEN-DESTEYGLFQISN  57
            K F +CEL+  +  + +D Y   +L   +C     S ++TQA   N D ST+YG+ QI++
Sbjct  1    KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINS  60

Query  58   ALWCKSSQSPQSRNICDITCDKFLDDDITDDIMCAKKIL-DIKGIDYWIAHKALC  111
              WC   ++P SRN+C+I C   L  DIT  + CAKKI+ D  G++ W+A +  C
Sbjct  61   RWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRC  115
```

How far hemoglobins in human and chicken are similar?

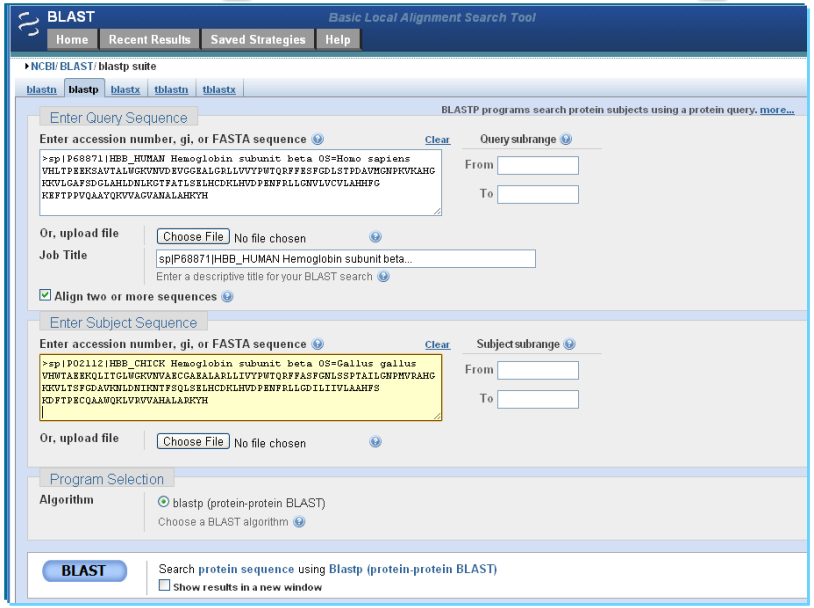
Get the sequences and compare the amino acids one by one

>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPD
AVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR
LLGNVLVCVLAHHFG KEFTPPVQAAYQKVVAGVANALAHKYH

>sp|P02112|HBB_CHICK Hemoglobin subunit beta OS=Gallus gallus
VHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSP
TAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENF
RLLGDILIIVLAAHFS KDFTPECQAAWQKLVRVVAHALARKYH

Comparing Human and Chicken protein sequences



Bioinformatics program BLAST

Give sequences

1. Human
2. Chicken

```
 Identities = 102/147 (69%), Positives = 121/147 (82%), Gaps = 0/147 (0%)

Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MVH T EEK  +T LWGKVNV E G EAL RLL+VYPWTQRFF SFG+LS+P A++GNP
Sbjct  1    MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM  60

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
            V+AHGKKVL +F D + +LDN+K TF+ LSELHCDKLHVDPENFRLLG++L+ VLA HF
Sbjct  61   VRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS  120

Query  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
            K+FTP  QAA+QK+V  VA+ALA KYH
Sbjct  121  KDFTPECQAAWQKLVRVVAHALARKYH  147
```

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- FASTA Help
- MView Help
- VisualFASTA Help

- View all FASTA's at EBI
- FASTA Programmatic Access

- Database Information
  - UniProt
  - UniParc

- Similar Applications
  - FASTA
  - BLAST
  - GGSEARCH
  - GLSEARCH
  - MPsrch
  - ScanPS
  - SSEARCH

FASTA Related Literature

Search for FASTA related literature in Medline...
more

EBI > Tools > Similarity & Homology

FASTA/SSEARCH/GGSEARCH/GLSEARCH - Protein Similarity Search

Provides sequence similarity searching against protein databases using the FASTA and SSEARCH programs. **SSEARCH** does a rigorous Smith-Waterman search for similarity between a query sequence and a database. **GGSEARCH** compares a protein or DNA sequence to a sequence database producing global-global alignment (Needleman-Wunsch). **GLSEARCH** compares a protein or DNA sequence to a sequence database. **FASTA** can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity searching against nucleotide databases or complete proteome/genome databases using the FASTA programs.

Download Software

PROGRAM	DATABASES	RESULTS	SEARCH TITLE	YOUR EMAIL
FASTA	Protein (UniProt Knowledgebas; UniProtKB/Swiss-Prot; UniProt Clusters 100%)	interacti	Sequence	

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM	-10	-2	2	10.0	default

DNA STRAND	HISTOGRAM	MOLECULE TYPE
none	no	Protein

SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	FILTER	STATISTICAL ESTIMATES
50	50	START-END	START-END	none	Regress

Enter or Paste a PROTEI Sequence in any format: Help

```
>2LZM:A|PDBID|CHAIN|SEQUENCE
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNC
NGVITKDEAEKLFNQDVDAAVRGILR
NAKLKPVYDSLDAVRRCALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLA
KSRWYNQTPNRAKRVITTFRTGTWDA
YKNL
```

Upload a file: Browse... Run Reset

FASTA Results

SUBMISSION PARAMETERS			
Title	Sequence	Database	uniprot
Sequence length	164	Sequence type	p
Program	fasta	Version	35.04 July. 4, 2009
Expectation upper value	10.0	Matrix	BL50
Sequence range	1-	Number of scores	50
Number of alignments	50	Word size	2
Open gap penalty	-10	Gap extension penalty	-2
Histogram	false		

Show Annotation | FASTA Result | MView | VisualFasta | XML | SUBMIT ANOTHER JOB

Clear all | Check all | Invert selection | Reset | Show Alignments | Download | fasta

Alignment	DB:ID	Source	Length	Identity%	Similar%	Overlap	E()
1	UNIPROT:LYS_BPT4	Lysozyme OS=Enterobacteria pha	164	100.0	100.0	164	2.3e-67
2	UNIPROT:C3V1I9_9CAUD	Soluble lysozyme OS=Entero	164	98.8	100.0	164	1.2e-66
3	UNIPROT:Q7M2A4_BPT2	Lysozyme OS=Enterobacteria	164	98.2	100.0	164	3.1e-66
4	UNIPROT:Q06EK2_BPR32	Lysozyme OS=Enterobacteria	164	98.2	100.0	164	3.1e-66
5	UNIPROT:C3V2B5_BPR51	Soluble lysozyme OS=Entero	164	98.2	100.0	164	3.1e-66
6	UNIPROT:Q7Y2B5_BPR69	Lysozyme OS=Enterobacteria	157	82.7	93.6	156	2.8e-53
7	UNIPROT:A8R9C2_9CAUD	Lysozyme OS=Enterobacteria	162	75.3	91.4	162	9.6e-50
8	UNIPROT:C4MZK9_9CAUD	E Lysozyme murein hydrolas	162	75.3	91.4	162	9.6e-50
9	UNIPROT:LYST1_DICDI	Probable T4-type lysozyme 1	170	46.7	76.6	167	1.1e-26
10	UNIPROT:LYST2_DICDI	Probable T4-type lysozyme 2	170	45.1	76.8	164	7.8e-25
11	UNIPROT:Q56EM5_9CAUD	Lysozyme OS=Aeromonas phag	600	42.9	70.6	163	2e-21
12	UNIPROT:Q6U9G4_9CAUD	Lysozyme OS=Aeromonas phag	600	42.9	71.2	163	2.4e-21
13	UNIPROT:Q19CF2_9CAUD	Lysozyme OS=Aeromonas phag	164	42.7	68.3	164	2.5e-21
14	UNIPROT:A7XF92_9CAUD	Lysozyme OS=Enterobacteria	599	43.2	69.8	162	1e-20
15	UNIPROT:C4MYV6_9CAUD	Baseplate hub subunit and	599	43.2	69.8	162	1e-20

Multiple sequence alignment

A **multiple sequence alignment** (MSA) is a sequence alignment of **three or more** biological sequences, generally **protein, DNA, or RNA**.

The input set of query sequences are assumed to have an **evolutionary relationship** by which they share a lineage and are descended from a **common ancestor**.

Multiple sequence alignment is a 2D table in which the rows represent individual sequences and the columns the residue positions.

Sequences are laid on the grid in such a way that (i) the relative positioning of residues within any sequence is preserved and (ii) similar residues in all sequences are brought into vertical register.

	1	2	3	4	5	6	7	8	9	10
I	Y	D	G	G	A	V	-	E	A	L
II	Y	D	G	G	-	-	-	E	A	L
III	F	E	G	G	I	L	V	E	A	L
IV	F	D	-	G	I	L	V	Q	A	V
V	Y	E	G	G	A	V	V	Q	A	L
Consensus	y	d	G	G	A/I	V/L	V	e	A	l

Computational complexity

Pairwise alignment techniques generally use processing time and memory space related to the products of the lengths of the sequences being compared [($O(m_1m_2)$, where O is the order of the time taken and m: sequence lengths).

By extending the pairwise comparison to three dimensions, we have the complexity of $O(m_1m_2m_3)$

For n sequences it will be $O(m^n)$, n is the number of sequences and m is the length of the sequences.

Time taken to compute an alignment rises exponentially with the number of mn sequences that are to be aligned.

Clustal

The Clustal approach exploits the fact that similar sequences are likely to be evolutionarily related.

Clustal aligns sequences in pairs, following the branching order of a family tree.

Similar sequences are aligned first, and more distantly related sequences are added later.

Once pairwise alignment scores for each sequences relative to all others have been calculated, they are used to cluster the sequences into groups, which are then aligned against each other to generate the final multiple alignment.

ClustalW uses the positioning of gaps in closely related sequences to guide the insertion of gaps into those that are more distant.

Information compiled during the alignment process about the variability of the most similar sequences is used to help vary the gap penalties on a residue and position specific basis.

ClustalW2

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

New users, please read the FAQ.

>> Download Software

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	
	Sequence	interactive	full	
KTUP (WORD SIZE)	**WINDOW LENGTH**	**SCORE TYPE**	**TOPDIAG**	**PAIRGAP**
def	def	percent	def	def
MATRIX	**GAP OPEN**	**NO END GAPS**	**GAP EXTENSION**	**GAP DISTANCES**
def	def	yes	def	def
	ITERATION		**NUMITER**	
	none		1	

OUTPUT		PHYLOGENETIC TREE			
OUTPUT FORMAT	**OUTPUT ORDER**	**TREE TYPE**	**CORRECT DIST.**	**IGNORE GAPS**	**CLUSTERING**
aln w/numbers	aligned	none	off	off	NJ

Enter or paste a set of sequences in any supported format: Help

```
>479227|Genbank|Outer membrane integral membrane
protein|outD protein
MLLLSGSVLLMASSLAWSAEFSASFKGTDIQEFINTVSKNLNKTVIIDPSVSGT
ITVRSY
DMMNEEQYYQFFLSVLDVYGFTVIPMDNNVLKIIRSKDAKSTSMPLATDRQPGI
GDEVVT
RVVPVNNVAARDFGRSSRVERQRVAWDVWRLRTCERRRDDWPRGVIHAVMTIVE
RVDQTG
DRNVTTIPLSYASSTEVVKMVNELNKMDEKSALPGMLTANVVADERTNSAAGFG
EPNSRQ
```

Upload a file: Browse... Run Reset

Sequence	Alignment	R_p
479227|Genbank|Outer	--FKGTDIQEFINTVS-KNLNKTVIIDPS-VSGTITVRSYDMMNEEQ---	67
P31701|SwissProt|Outer	--FKGTDIQEFINTVS-KNLNKTVIIDPS-VSGTITVRSYDMMNEEQ---	67
P31700|SwissProt|Outer	--FKGTDIQEFINTVS-KNLNKTVIIDPT-VRGTISVRSYDMMDEGQ---	76
Q01565|SwissProt|Outer	--FKGTDIQEFINTVS-KNLNKTVIIDPT-VRGTISVRSYDMMNEGQ---	76
7466966|Genbank|Outer	--FKDTDIQEFINTVS-KNLHKTVIINPD-VQGTITVRSYDMLNEEQ---	63
P15644|SwissProt|Outer	--FKGTDIQEFINTVS-KNLNKTVIIDPS-VRGTITVRSYDMLNEEQ---	76
P31780|SwissProt|Outer	--FKNADIEEFINTVG-KNLSKTIIIEPS-VRGKINVRSYDLLNEEQ---	74
P45778|SwissProt|Outer	--FKNADIEEFINTVG-KNLSKTIIIEPS-VRGKINVRSYDLLNEEQ---	74
P45779|SwissProt|Outer	--FKGTDIQEFINIVG-RNLEKTIIVDPS-VRGKVDVRSFDTLNEEQ---	73
4139236|Genbank|Outer	--FVNADIDQVAKAIG-AATGKTIIVDPR-VKGQLNLVAERPVPEDQ---	69
11559475|Genbank|Outer	--FVNADIDQVAKAIG-AATGKTIIVDPR-VKGQLNLVAERPVPEDQ---	76
3978475|Genbank|Outer	INMKDADIRDFIDQVA-QISGETFVVDPR-VKGQVTVISKTPLGLEE---	91
P35818|SwissProt|Outer	INLKDADIREFIDQIS-EITGETFVVDPR-VKGQVSVVSKAQLSLSE---	99
11352555|Genbank|Outer	INMKDAEIGDFIEQVS-SISGQTFVVDPR-VKGRVTVVSQARLSLAE---	92
15597065|Genbank|Outer	INMKDAEIGDFIEQVS-SISGQTFVVDPR-VKGRVTVVSQARLSLAE---	92
2120685|Genbank|Outer	VNFVDTELGEFIDSVS-RITGTTFIVDPR-VKGKVTVRTVDLHDADA---	83
13421292|Genbank|Outer	LNVQDADIRVFIQDVA-KSTGTTFIIDPR-VKGTVTVASNGPLNRRE---	81
7469078|Genbank|Outer	--FEDISILELLQFVS-KISGTNFVFDSNDLQFNVTIVSHDPTSVDD---	312
11362809|Genbank|Outer	--FEDISILELLQFVS-KISGTNFVFDSNDLQFNVTIVSHDPTSVDD---	63
7468594|Genbank|Outer	--FEDISILELLQFVS-KISGTNFVFDSNDLQFNVTIVSHDPTSVDD---	310
11360974|Genbank|Outer	FNFEGESLQAVVKAILGDMLGQNYFIASG-VQGTVTLSTPKPVSSAQ---	153
P29041|SwissProt|Outer	FNFEGESVQAVVKAILGDMLGQNYVIAPG-VQGTVTLATPNPVSPAQ---	141
13475694|Genbank|Outer	LNLVNAPIADAAKAVLGDALHLNYIVDPR-VQGTVTLQTSQPVSQDA---	139
12721580|Genbank|Outer	----------MWRAFRKIS-LVYFLCGVAYVGS---	22
P31772|SwissProt|Outer	----------MKKYFLKCGYFLVCFCLPLIVFA---	23
11354911|Genbank|Outer	-ALCASSMVFSAESATANQLENIDFRVNKEKAAVLIVELASPSAVVD---	73
P34749|SwissProt|Outer	----------MKQWIAALLLMLIPGVQAAKP---	21
11360960|Genbank|Outer	-AENKQAIPVPKVANAPLSVSKIDFKRGDDGSGRLILKFDGQGATPD---	93
P34750|SwissProt|Outer	-ASYAQPIKPKPYVPAGRAIRNIDFQRGEKGEGNVVIDLSDPTLSPD---	190
4027986|Genbank|Outer	SAKQQAAAPAKQQAAAPAKQTNIDFRKDGKNAGIIELAALGFAGQPD---	260
11353851|Genbank|Outer	SAKQQTAAPAKQQAATPAKQTNIDFRKDGKNAGIIELAALGFAGQPD---	244
4768955|Genbank|Outer	----------	
12721161|Genbank|Outer	----------	
7208425|Genbank|Outer	VNLPSVKASMSASRRLLTASVAALLALTSTAPVFADGPIGGSHTYRP---	51
13474660|Genbank|Outer	----------	
12620550|Genbank|Outer	--MKVLNNAAGATQPPAITNPQRSAALNRLCYLLCG----------	34
13475294|Genbank|Outer	----------	
7468922|Genbank|Outer	ALWNHPETTIYNLVSDYGDEQSIYVIPQNVGAMRITAMSKLVVPKEG---	164
11360973|Genbank|Outer	ALWNHPETTIYNLVSDYGDEQSIYLIPQNVGAMRITAMSKLVVPKEG---	164
7468239|Genbank|Outer	ALWNHPETTIYNLVTDYGTEDSIYLIPQEIGAIKIATLSKFVVPKES---	165

Exercise

Get the multiple alignment for hemoglobin A chain from different organisms.

Steps

1. Search for hemoglobin A chain sequences
2. Select relevant entries
3. Get the sequences in FASTA format (view or save).
4. Give the sequences as input for ClustalW

Other software

MAFFT

http://mafft.cbrc.jp/alignment/software/

MUSCLE

http://www.ebi.ac.uk/Tools/msa/muscle/

PROMOLS

http://prodata.swmed.edu/promals/promals.php

PSI-BLAST

PSI-BLAST (Position-Specific Iterative BLAST) is a program that searches database of sequences similar to query sequence.

PSI-BLAST begins with search results obtained with BLAST and derives pattern information from a multiple sequence alignment of the initial hits.

It repeats the process and fine-tuning the pattern in successive cycles.

PSI-BLAST

PSI-BLAST - Protein Similarity Search

PSI-BLAST is similar to NCBI BLAST except that it uses position-specific scoring matrices derived during the search, this tool is used to detect distant evolutionary relationships.

Use this tool

STEP 1 - Select your database

PROTEIN DATABASES

UniProt Knowledgebase

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1
PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
```

Upload a file: Choose File No file chosen

STEP 3 - Set your parameters

PSI-BLAST THRESHOLD

1.0e-3

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 4 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

PSI-BLAST Results

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details | Submit Another Job

PSI-BLAST Iteration

Threshold: 1.0e-3 Run next iteration

Alignments

Selection: Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

Download in fasta format

Clear Selection | Select All | Invert Selection

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
1 New	TR:Q5R9M5_PONAB	Putative uncharacterized protein DKFZp468J1717 OS=Pongo abelii GN=DKFZp468J1717 PE=2 SV=1 *Cross-references and related information in:* ▶Nucleotide Sequences ▶Genomes ▶Ontologies ▶Protein Families ▶Protein Sequences	142	733	100.0	100.0	5.0E-76
2 New	TR:D1MGQ2_HUMAN	Alpha-2 globin chain OS=Homo sapiens GN=HBA2 PE=3 SV=1 *Cross-references and related information in:* ▶Nucleotide Sequences ▶Genomes ▶Ontologies ▶Protein Families ▶Literature ▶Protein Sequences	142	733	100.0	100.0	5.0E-76
3 New	SP:HBA_PANTR	Hemoglobin subunit alpha OS=Pan troglodytes GN=HBA1 PE=1 SV=2 *Cross-references and related information in:* ▶Nucleotide Sequences ▶Ontologies ▶Protein Families ▶Literature ▶Protein Sequences	142	733	100.0	100.0	5.0E-76
4 New	SP:HBA_PANPA	Hemoglobin subunit alpha OS=Pan paniscus GN=HBA1 PE=1 SV=2 *Cross-references and related information in:* ▶Ontologies ▶Protein Families ▶Literature ▶Protein Sequences	142	733	100.0	100.0	5.0E-76