Hydrophobicity profile

Hydrophobicity profile is simply the plot of the hydrophobicity indices of the residues against their sequence numbers.

E.g. SAMPLEDATAWITHHYDINDICES

-1, 1, 1, -1, 2, -2, -2, 1, -1, 1, 2, 2, -1, -2, -2, 1, -2, 2, -1, -2, 2, 1, -2, -1

>1a91_
A MENLNMDLLY MAAAVMMGLA AIGAAIGIGI
LGGKFLEGAA RQPDLIPLLR TQFFIVMGLV
DAIPMIAVGL GLYVMFAVA

A, C, G, M, Y: 1
F, I, L, V, W: 2
D, E, H, K, R: -2
N, P, Q, S, T: -1

Sample data

Nozaki-Tanford-Jones (Ht)

A: 0.87 D: 0.66 C:1.52 E: 0.67

F: 2.87 G: 0.10 H: 0.87 I: 3.15

K: 1.64 L: 2.17 M: 1.67 N: 0.09

P: 2.77 Q: .00 R: 0.85 S: 0.07

T: 0.07 V: 1.87 W: 3.77 Y: 2.67

Ponnuswamy-Gromiha (Hgm)

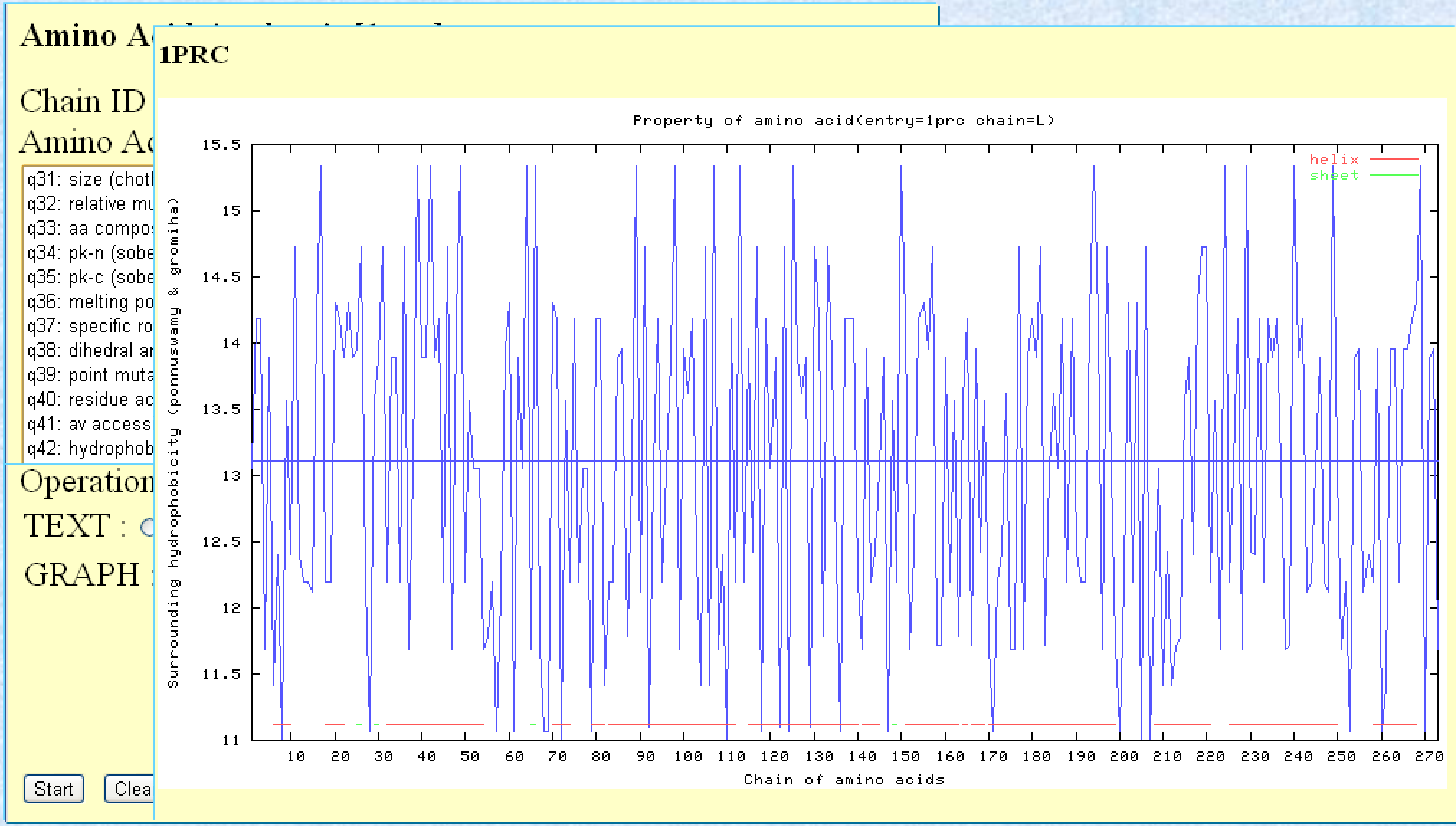
A: 13.85 D: 11.61 C: 15.37 E: 11.38

F: 13.93 G: 13.34 H: 13.82 I: 15.28

K: 11.58 L: 14.13 M: 13.86 N: 13.02

P: 12.35 Q: 12.61 R: 13.10 S: 13.39

T: 12.70 V: 14.56 W: 15.48 Y: 13.88



Amino A

Chain ID

Amino Ac

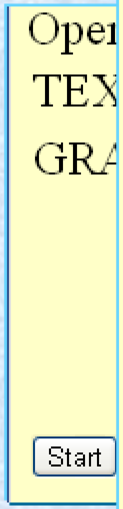
q31: size (chotl
q32: relative mu
q33: aa compos
q34: pk-n (sobe
q35: pk-c (sobe
q36: melting po
q37: specific ro
q38: dihedral ar
q39: point muta
q40: residue ac
q41: av access
q42: hydrophob

Operation

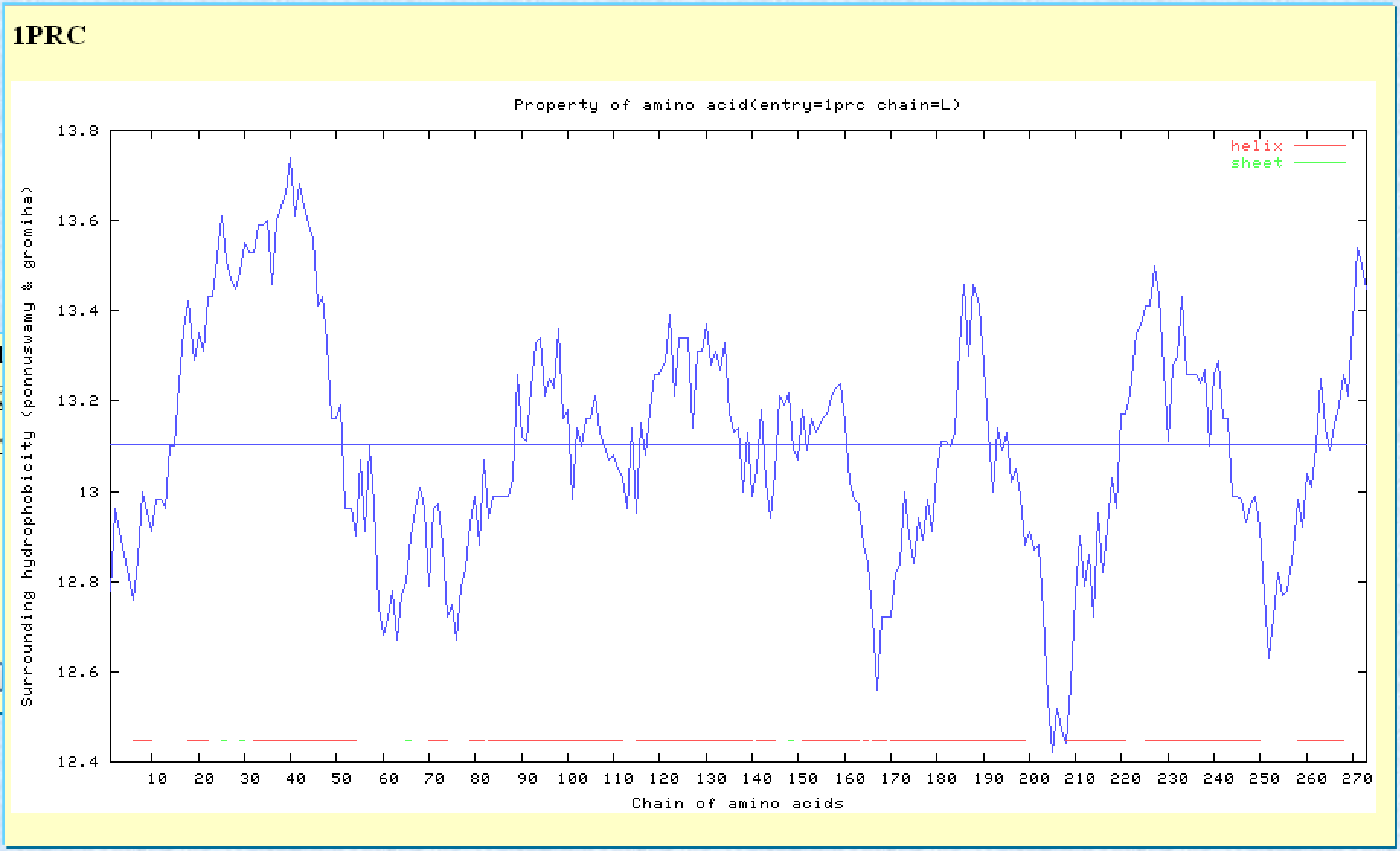
TEXT : ○

GRAPH :

1PRC



Oper
TEX
GRA
Start



1PRC
Property of amino acid(entry=1prc chain=L)
Surrounding hydrophobicity (ponnuswamy & gromiha)
13.8
13.6
13.4
13.2
13
12.8
12.6
12.4
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
Chain of amino acids
helix
sheet

Amphipathicity

Amphipathic character of amino acid residues is the periodicities in the ploar/nonpolar character of the amino acid sequence in a protein.

This has been examined by assigning a numerical hydrophobicity to each residue and searching for periodicity in the resulting one-dimensional function.

Amphipathicity: α-helices

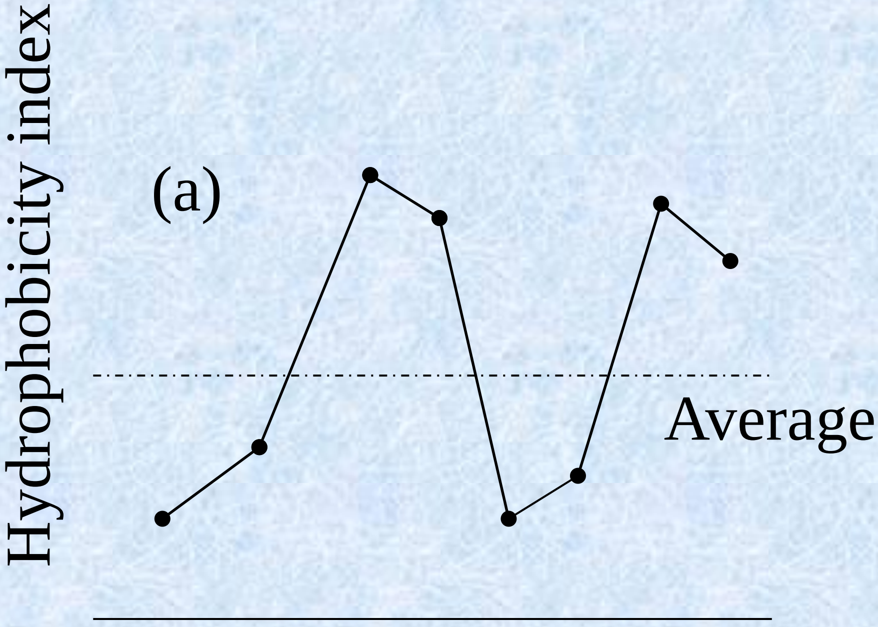
The residues of an α-helical segment are considered on four adjacent edges along the direction of the helical axis. The average hydrophobicity of the residues constituting the edge i (i = 1,4) is given by

$$\alpha_i = (\Sigma h_{i+j})/n,$$

where n is the total number of residues in the edge,

j increases at an interval of 4 from 0 to m, m being the number of residues in the helix;

h is the hydrophobic index of the residue.



The power of amphipathicity of a helix is taken to be

$$A_\alpha = |(a_1+a_2) - (a_3+a_4)| \text{ or } |(a_1+a_4) - (a_2+a_3)|.$$

It has been reported that 75% of the helical segments in known structures are amphipathic in nature.

Amphipathicity: β-strands

A β-strand segment is considered to have two faces and the average hydrophobicity of residues constituting the face i (i = 1, 2) is given by

$$\beta_i = (\Sigma h_{i+j})/n,$$

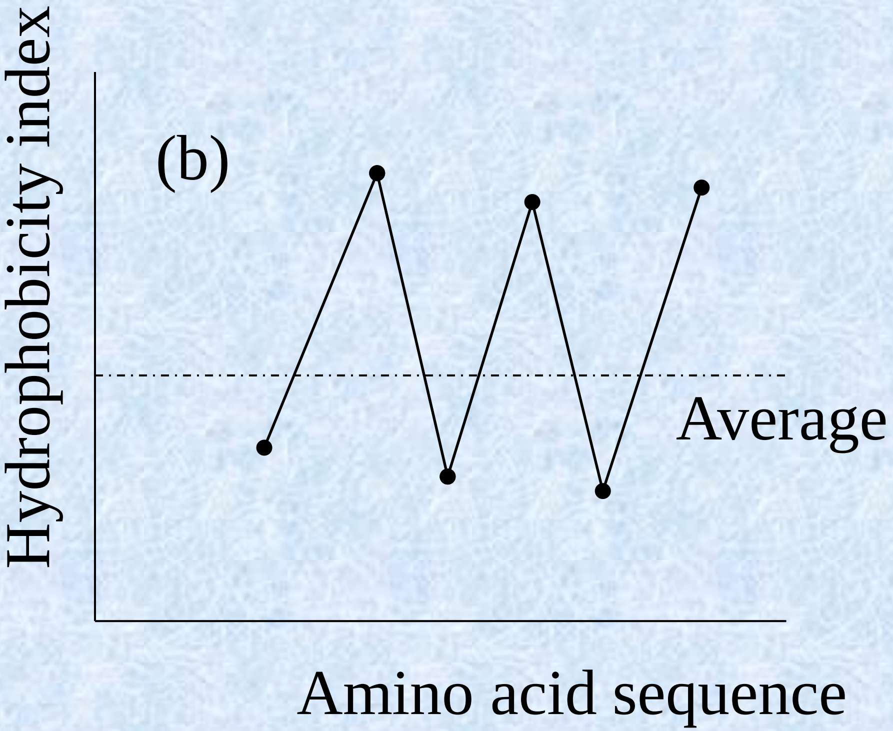
where n is the total number of residues in the face, j increases at an interval of 2 from 0 to m, m being the number of residues in the strand;

The amphipathicity index of a strand is computed using the equation,

$$A_\beta = |\beta_1 - \beta_2|.$$

The structural analysis showed that about 65% of the β-strands possess amphipathic character.

"



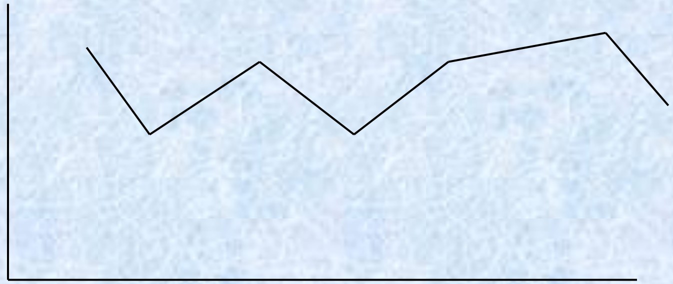
Patterns

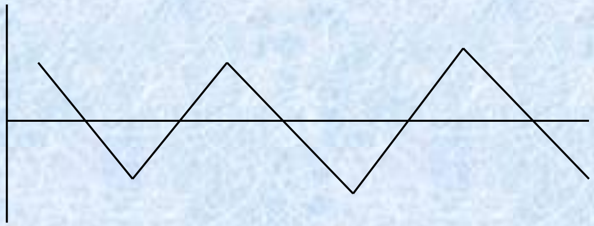
- Identify the pattern of hydrophobic residues for membrane spanning helical proteins

- Amphiphathic character of β-strands by alternative hydrophobic-hydrophilic residues

E.g. AILVGYWFFVVA

AKINIHVTFKIKLP





Pattern Definition

[LIVM]-[VIC]-x(2) -G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x (2)-G

1. Use capital letters for amino acid residues

2. Use "[…]" for a choice of multiple amino acids in a particular position. [LIVM] means that L, I, V, or M can be in the first position

3. Use "{…}" to exclude amino acids. {CF} means C and F should not be in that particular position

4. Use "x" or "X" for a position that can be any amino acid.

5. Use "(n)", where n is a number, for multiple positions; x(3) is the same as "xxx"

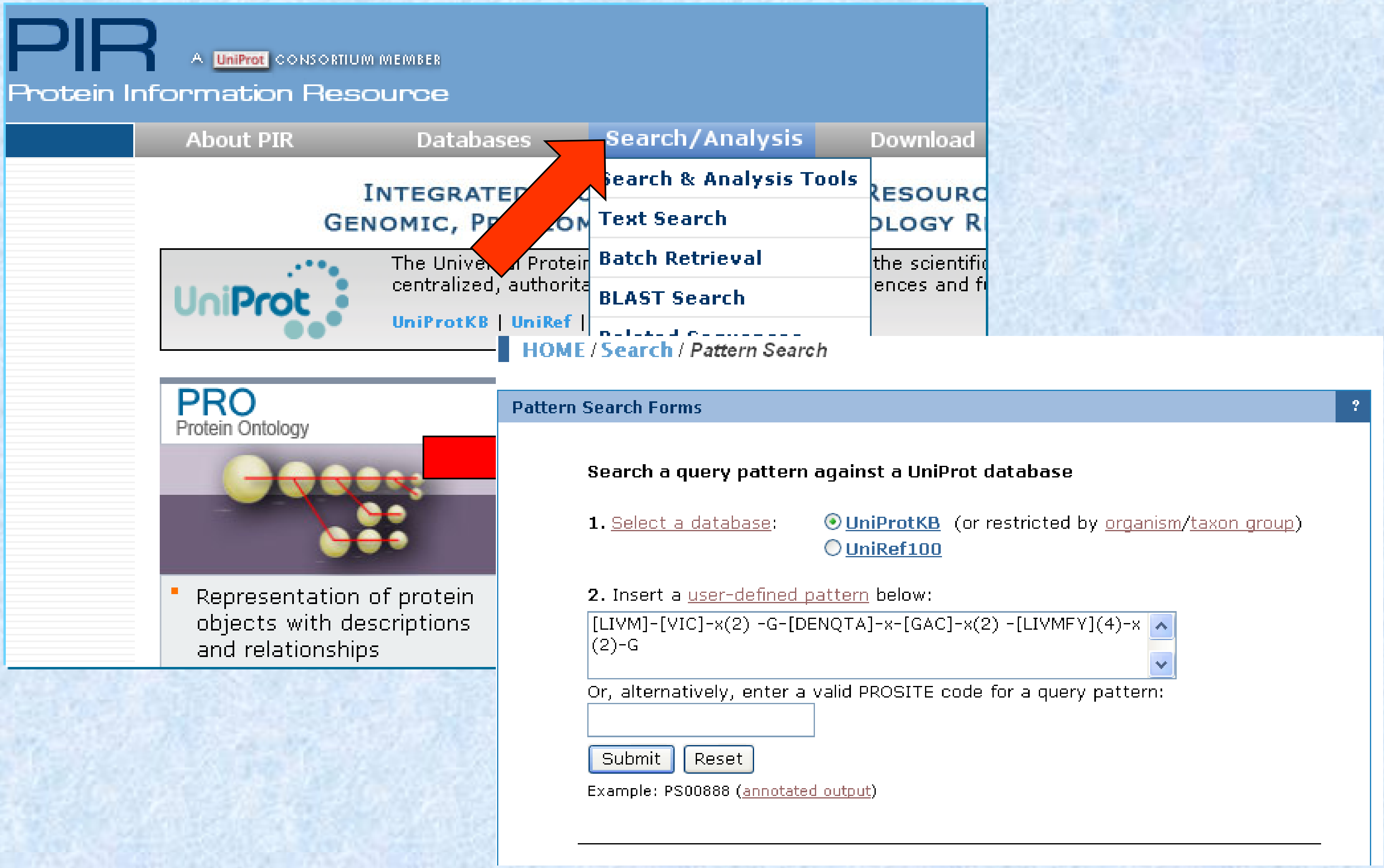
PIR: Pattern Definition

[LIVM]-[VIC]-x(2) -G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x (2)-G

Illustrates a 17 amino acid peptide that has:

L, I, V, or M at position 1;
V, I, or C at position 2;
any residue at positions 3 and 4;
G at position 5 and so on ….

PIR: Pattern search



HOME / Search / Pattern Search

Pattern Search Forms ?

Search a query pattern against a UniProt database

1. Select a database: ● UniProtKB (or restricted by organism/taxon group)
○ UniRef100

2. Insert a user-defined pattern below:

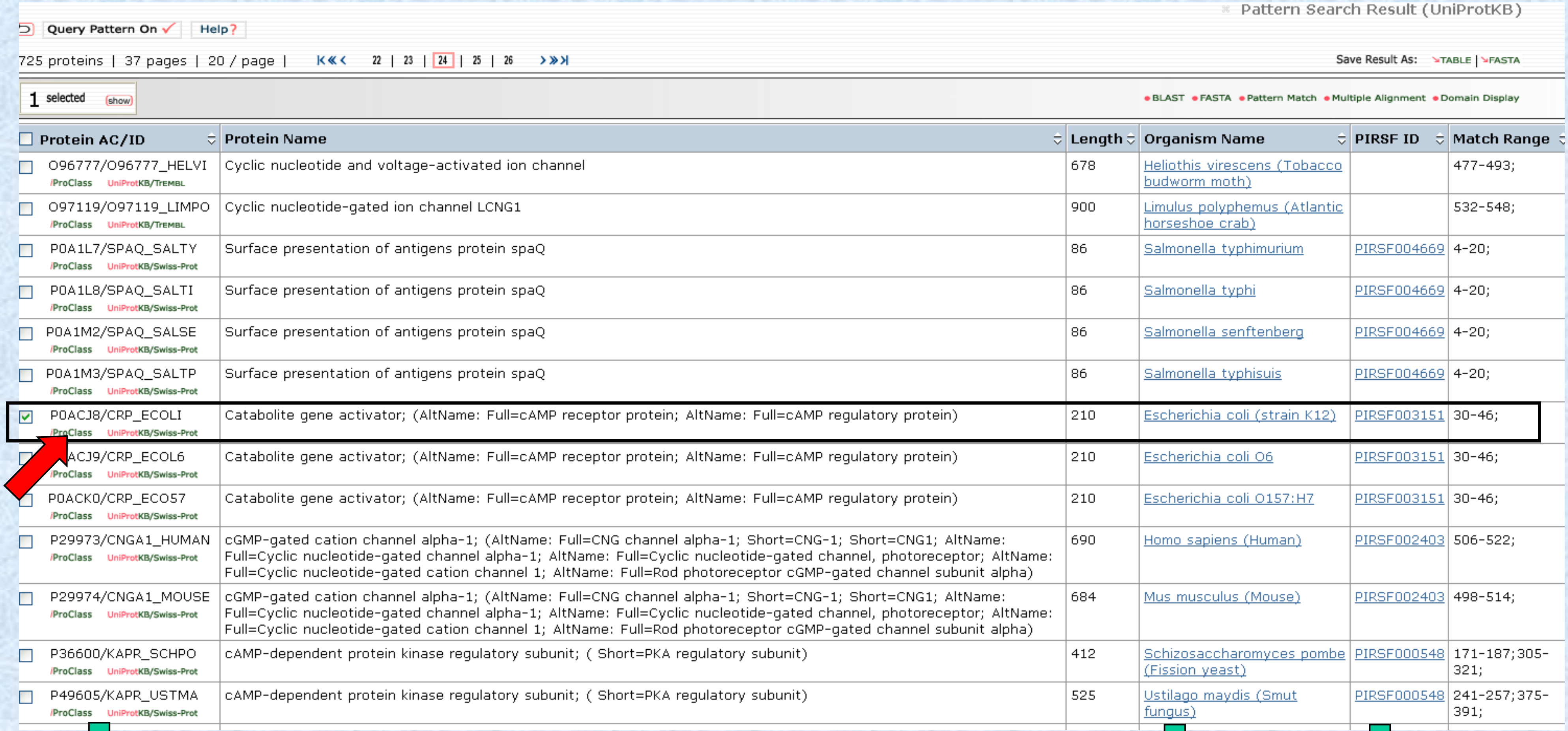
```
[LIVM]-[VIC]-x(2) -G-[DENQTA]-x-[GAC]-x(2) -[LIVMFY](4)-x
(2)-G
```

Or, alternatively, enter a valid PROSITE code for a query pattern:

Submit | Reset

Example: PS00888 (annotated output)

[LIVM]-[VIC]-x(2) -G-[DENQTA]-x-[GAC]-x(2) -[LIVMFY](4)-x(2)-G



Pattern Search Result (UniProtKB)

Query Pattern On ✓ | Help?

725 proteins | 37 pages | 20 / page | K«‹ 22 | 23 | 24 | 25 | 26 ›»K

Save Result As: TABLE | FASTA

1 selected (show)

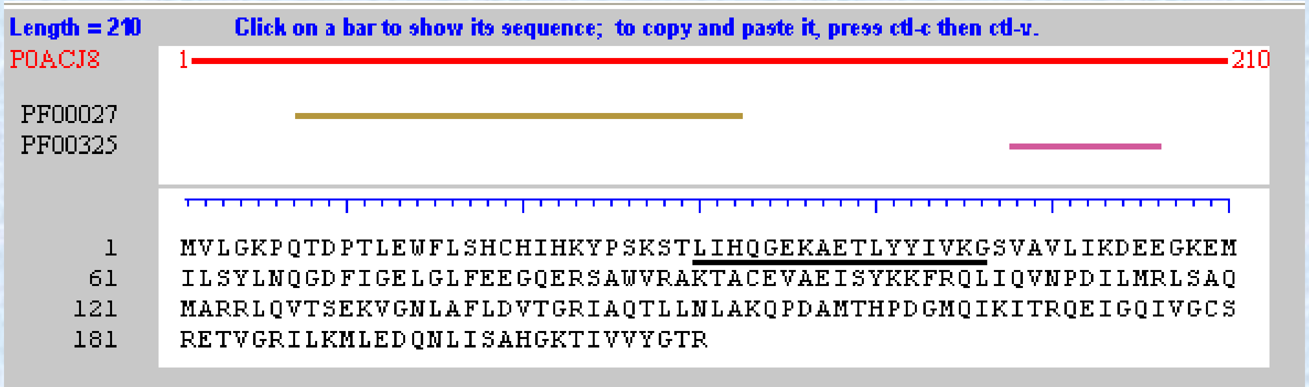
● BLAST ● FASTA ● Pattern Match ● Multiple Alignment ● Domain Display

☐ Protein AC/ID	Protein Name	Length	Organism Name	PIRSF ID	Match Range
☐ O96777/O96777_HELVI iProClass UniProtKB/TrEMBL	Cyclic nucleotide and voltage-activated ion channel	678	Heliothis virescens (Tobacco budworm moth)		477-493;
☐ O97119/O97119_LIMPO iProClass UniProtKB/TrEMBL	Cyclic nucleotide-gated ion channel LCNG1	900	Limulus polyphemus (Atlantic horseshoe crab)		532-548;
☐ P0A1L7/SPAQ_SALTY iProClass UniProtKB/Swiss-Prot	Surface presentation of antigens protein spaQ	86	Salmonella typhimurium	PIRSF004669	4-20;
☐ P0A1L8/SPAQ_SALTI iProClass UniProtKB/Swiss-Prot	Surface presentation of antigens protein spaQ	86	Salmonella typhi	PIRSF004669	4-20;
☐ P0A1M2/SPAQ_SALSE iProClass UniProtKB/Swiss-Prot	Surface presentation of antigens protein spaQ	86	Salmonella senftenberg	PIRSF004669	4-20;
☐ P0A1M3/SPAQ_SALTP iProClass UniProtKB/Swiss-Prot	Surface presentation of antigens protein spaQ	86	Salmonella typhisuis	PIRSF004669	4-20;
☑ P0ACJ8/CRP_ECOLI iProClass UniProtKB/Swiss-Prot	Catabolite gene activator; (AltName: Full=cAMP receptor protein; AltName: Full=cAMP regulatory protein)	210	Escherichia coli (strain K12)	PIRSF003151	30-46;
☐ P0ACJ9/CRP_ECOL6 iProClass UniProtKB/Swiss-Prot	Catabolite gene activator; (AltName: Full=cAMP receptor protein; AltName: Full=cAMP regulatory protein)	210	Escherichia coli O6	PIRSF003151	30-46;
☐ P0ACK0/CRP_ECO57 iProClass UniProtKB/Swiss-Prot	Catabolite gene activator; (AltName: Full=cAMP receptor protein; AltName: Full=cAMP regulatory protein)	210	Escherichia coli O157:H7	PIRSF003151	30-46;
☐ P29973/CNGA1_HUMAN iProClass UniProtKB/Swiss-Prot	cGMP-gated cation channel alpha-1; (AltName: Full=CNG channel alpha-1; Short=CNG-1; Short=CNG1; AltName: Full=Cyclic nucleotide-gated channel alpha-1; AltName: Full=Cyclic nucleotide-gated channel, photoreceptor; AltName: Full=Cyclic nucleotide-gated cation channel 1; AltName: Full=Rod photoreceptor cGMP-gated channel subunit alpha)	690	Homo sapiens (Human)	PIRSF002403	506-522;
☐ P29974/CNGA1_MOUSE iProClass UniProtKB/Swiss-Prot	cGMP-gated cation channel alpha-1; (AltName: Full=CNG channel alpha-1; Short=CNG-1; Short=CNG1; AltName: Full=Cyclic nucleotide-gated channel alpha-1; AltName: Full=Cyclic nucleotide-gated channel, photoreceptor; AltName: Full=Cyclic nucleotide-gated cation channel 1; AltName: Full=Rod photoreceptor cGMP-gated channel subunit alpha)	684	Mus musculus (Mouse)	PIRSF002403	498-514;
☐ P36600/KAPR_SCHPO iProClass UniProtKB/Swiss-Prot	cAMP-dependent protein kinase regulatory subunit; (Short=PKA regulatory subunit)	412	Schizosaccharomyces pombe (Fission yeast)	PIRSF000548	171-187;305-321;
☐ P49605/KAPR_USTMA iProClass UniProtKB/Swiss-Prot	cAMP-dependent protein kinase regulatory subunit; (Short=PKA regulatory subunit)	525	Ustilago maydis (Smut fungus)	PIRSF000548	241-257;375-391;

Links to iProClass and UniProtKB

Link to NCBI taxonomy

Link to PIRSF report


Length = 210
Click on a bar to show its sequence; to copy and paste it, press ctl-c then ctl-v.
P0ACJ8
1
210
PF00027
PF00325

```
  1   MVLGKPQTDPTLEWFLSHCHIHKYPSKSTLIHQGEKAETLYYIVKGSVAVLIKDEEGKEM
 61   ILSYLNQGDFIGELGLFEEGQERSAWVRAKTACEVAEISYKKFRQLIQVNPDILMRLSAQ
121   MARRLQVTSEKVGNLAFLDVTGRIAQTLLNLAKQPDAMTHPDGMQIKITRQEIGQIVGCS
181   RETVGRILKMLEDQNLISAHGKTIVVYGTR
```

Pattern: LIHQGEKAETLYYIVKG

[LIVM]-[VIC]-x(2) -G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x (2)-G

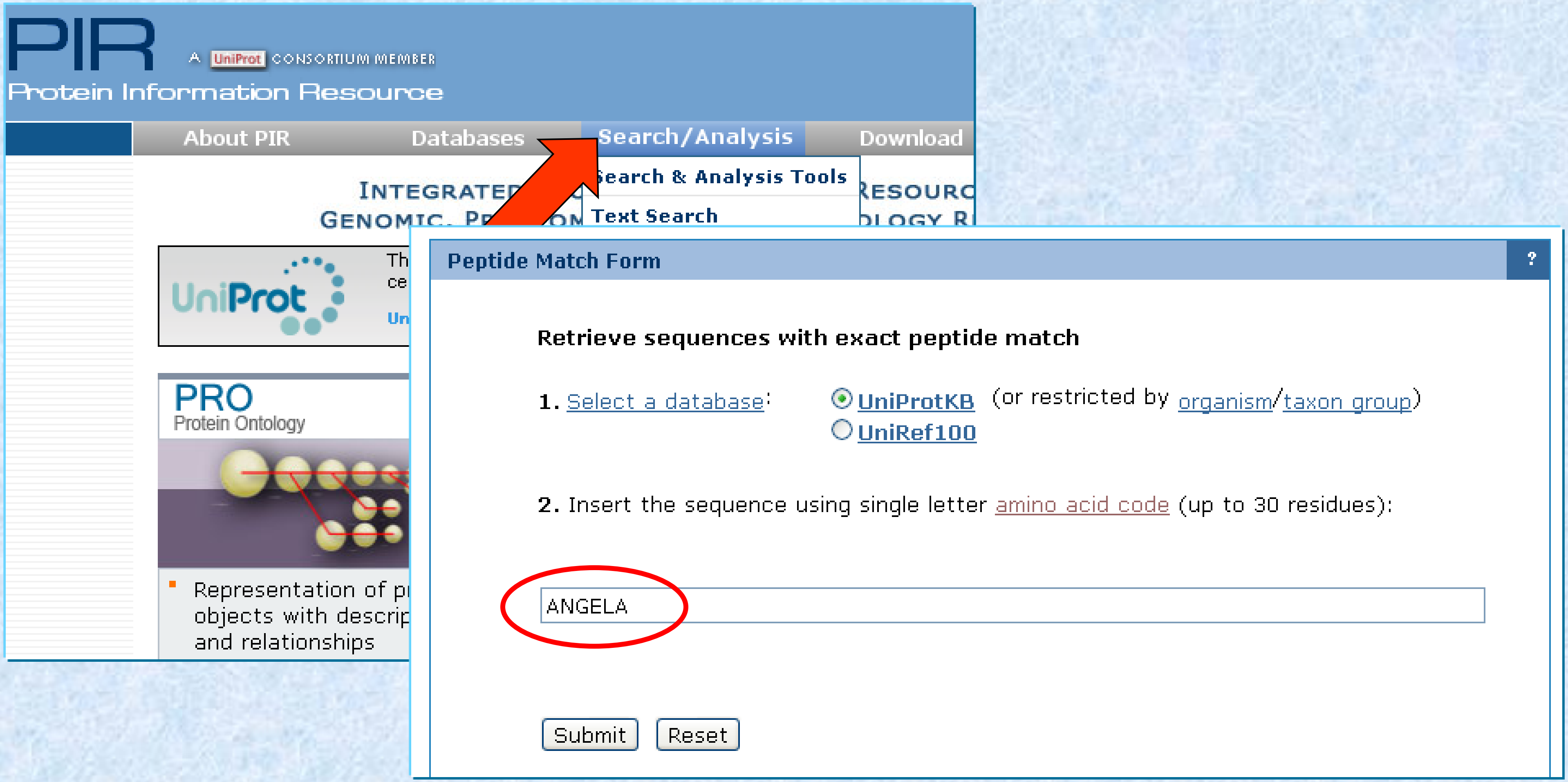
New β-signal motif : $P_{o}xGh_{y}xH_{y}xH_{y}$

[K,R,H,Q,N,S,T].G [I,V,L,F,M,Y,W,A,C].[I,V,L,F,M,Y,W].[I,V,L,F,M,Y,W]

Algorithm

K. Imai, M.M. Gromiha and P. Horton (2008) Cell

PIR: Specific Peptide search



Peptide Match Form ?

Retrieve sequences with exact peptide match

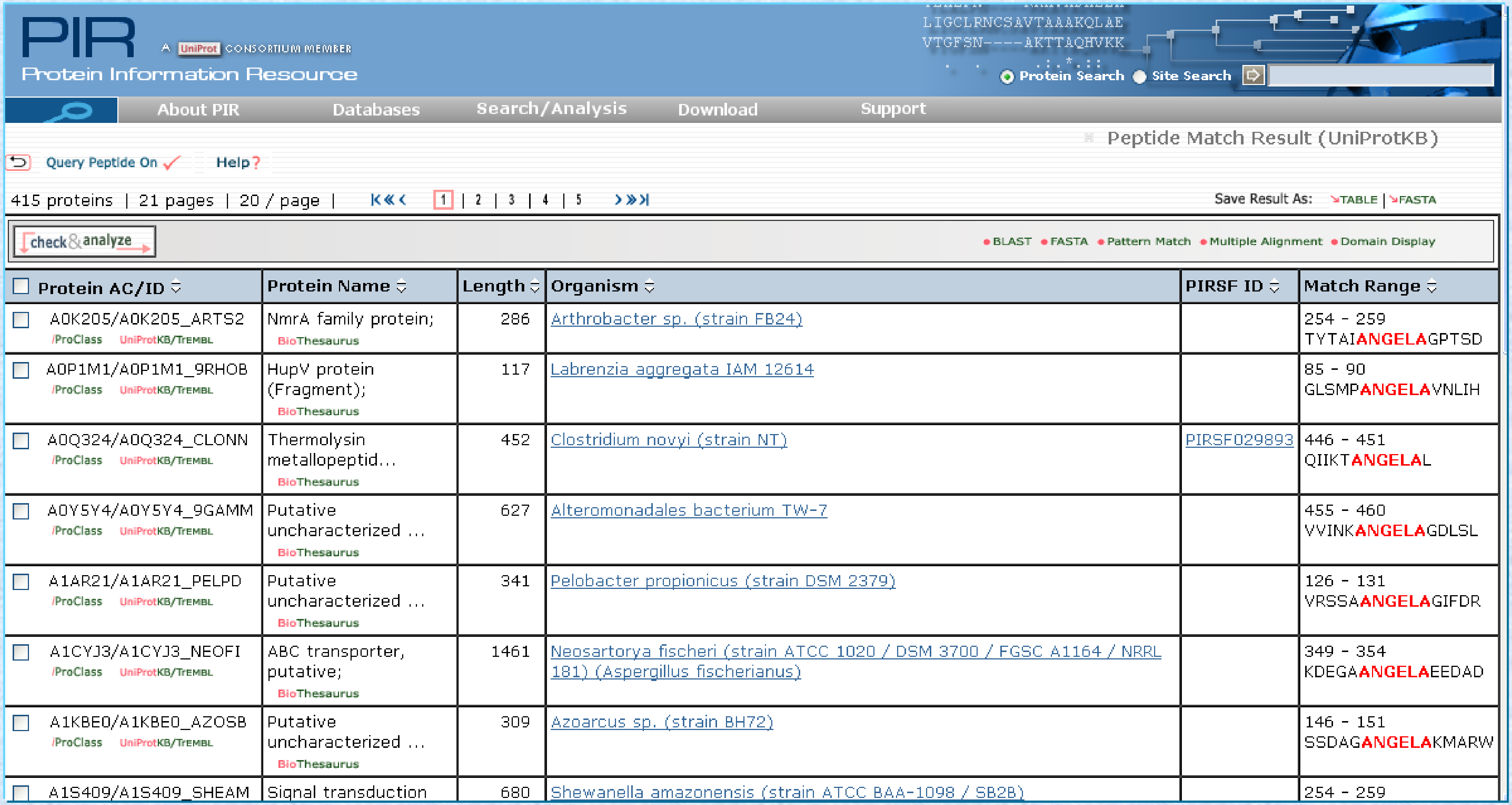
1. Select a database: ● UniProtKB (or restricted by organism/taxon group)
○ UniRef100

2. Insert the sequence using single letter amino acid code (up to 30 residues):

ANGELA

Submit | Reset

PIR: Specific Peptide search



Peptide Match Result (UniProtKB)

415 proteins | 21 pages | 20 / page | |◀◀ ◀ 1 | 2 | 3 | 4 | 5 ▶ ▶▶|

Save Result As: TABLE | FASTA

□ Protein AC/ID	Protein Name	Length	Organism	PIRSF ID	Match Range
□ A0K205/A0K205_ARTS2 ProClass UniProtKB/TrEMBL	NmrA family protein; BioThesaurus	286	Arthrobacter sp. (strain FB24)		254 - 259 TYTAIANGELAGPTSD
□ A0P1M1/A0P1M1_9RHOB ProClass UniProtKB/TrEMBL	HupV protein (Fragment); BioThesaurus	117	Labrenzia aggregata IAM 12614		85 - 90 GLSMPANGELAVNLIH
□ A0Q324/A0Q324_CLONN ProClass UniProtKB/TrEMBL	Thermolysin metallopeptid... BioThesaurus	452	Clostridium novyi (strain NT)	PIRSF029893	446 - 451 QIIKTANGELAL
□ A0Y5Y4/A0Y5Y4_9GAMM ProClass UniProtKB/TrEMBL	Putative uncharacterized ... BioThesaurus	627	Alteromonadales bacterium TW-7		455 - 460 VVINKANGELAGDLSL
□ A1AR21/A1AR21_PELPD ProClass UniProtKB/TrEMBL	Putative uncharacterized ... BioThesaurus	341	Pelobacter propionicus (strain DSM 2379)		126 - 131 VRSSAANGELAGIFDR
□ A1CYJ3/A1CYJ3_NEOFI ProClass UniProtKB/TrEMBL	ABC transporter, putative; BioThesaurus	1461	Neosartorya fischeri (strain ATCC 1020 / DSM 3700 / FGSC A1164 / NRRL 181) (Aspergillus fischerianus)		349 - 354 KDEGAANGELAEEDAD
□ A1KBE0/A1KBE0_AZOSB ProClass UniProtKB/TrEMBL	Putative uncharacterized ... BioThesaurus	309	Azoarcus sp. (strain BH72)		146 - 151 SSDAGANGELAKMARW
□ A1S409/A1S409_SHEAM	Signal transduction	680	Shewanella amazonensis (strain ATCC BAA-1098 / SB2B)		254 - 259

Position specific scoring matrices (Profiles)

Position specific scoring matrices (PSSM) or profiles express the patterns inherent in a multiple sequence alignment of a set of homologous sequences.

The basic idea to use profiles is to match the query sequences from the database against the sequences in the alignment table, giving higher weight to positions that are conserved than to those are variable.

These profiles are obtained with a set of probability scores for each amino acid (or gap) at each position of the alignment.

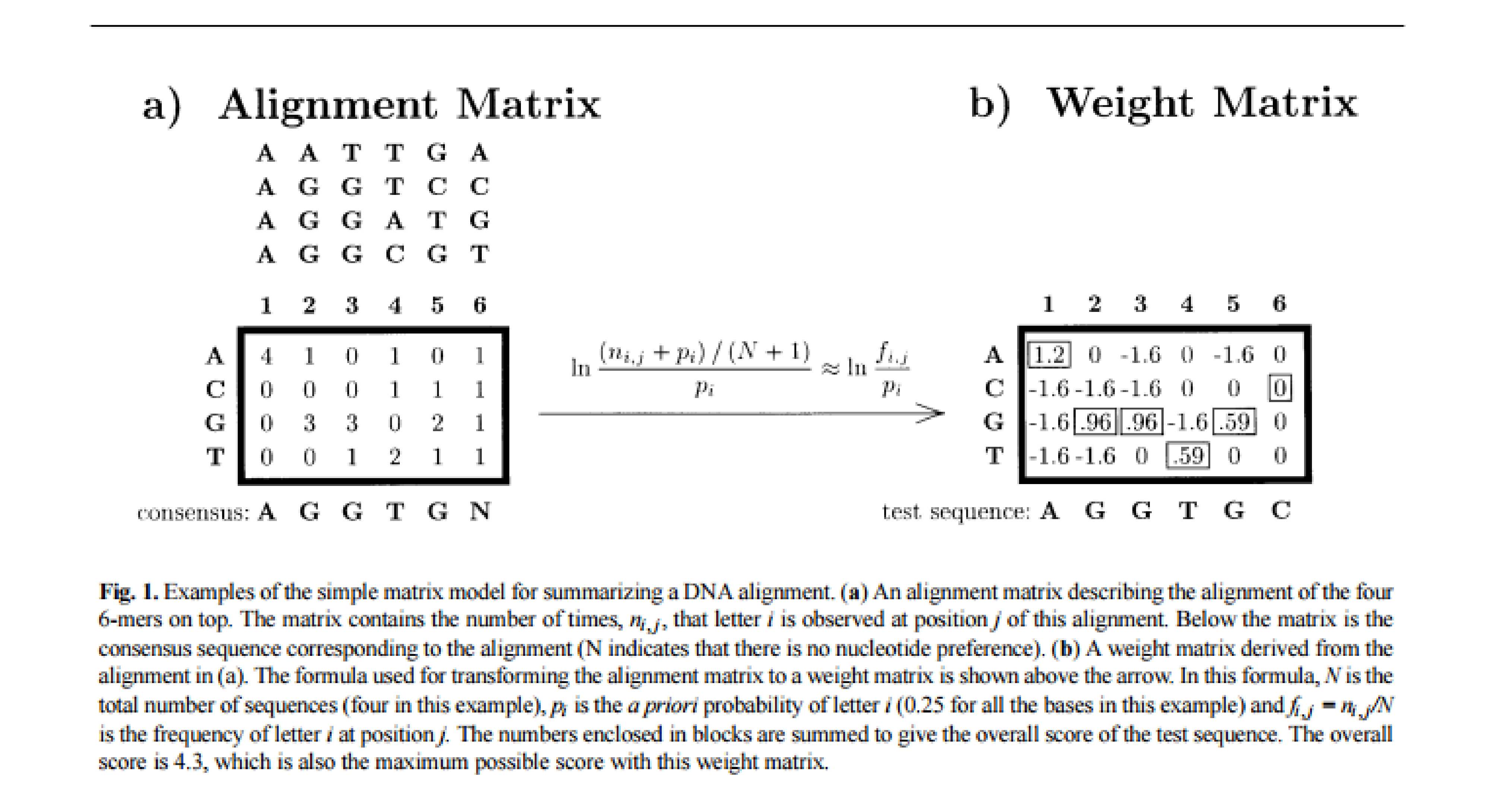
Profiles: Applications

(i) they permit greater accuracy in alignments of distantly-related sequences,

(ii) the conservation patterns facilitate identification of other homologous sequences,

(iii) patterns from the sequences are useful in classifying subfamilies within a set of homologues,

(iv) most structure prediction methods are reliable if based on multiple sequence alignment rather than on a single sequence etc.



a) Alignment Matrix

```
A A T T G A
A G G T C C
A G G A T G
A G G C G T
```

	1	2	3	4	5	6
A	4	1	0	1	0	1
C	0	0	0	1	1	1
G	0	3	3	0	2	1
T	0	0	1	2	1	1

consensus: A G G T G N

$$\xrightarrow{\ln \frac{(n_{i,j}+p_i)\,/\,(N+1)}{p_i} \approx \ln \frac{f_{i,j}}{p_i}}$$

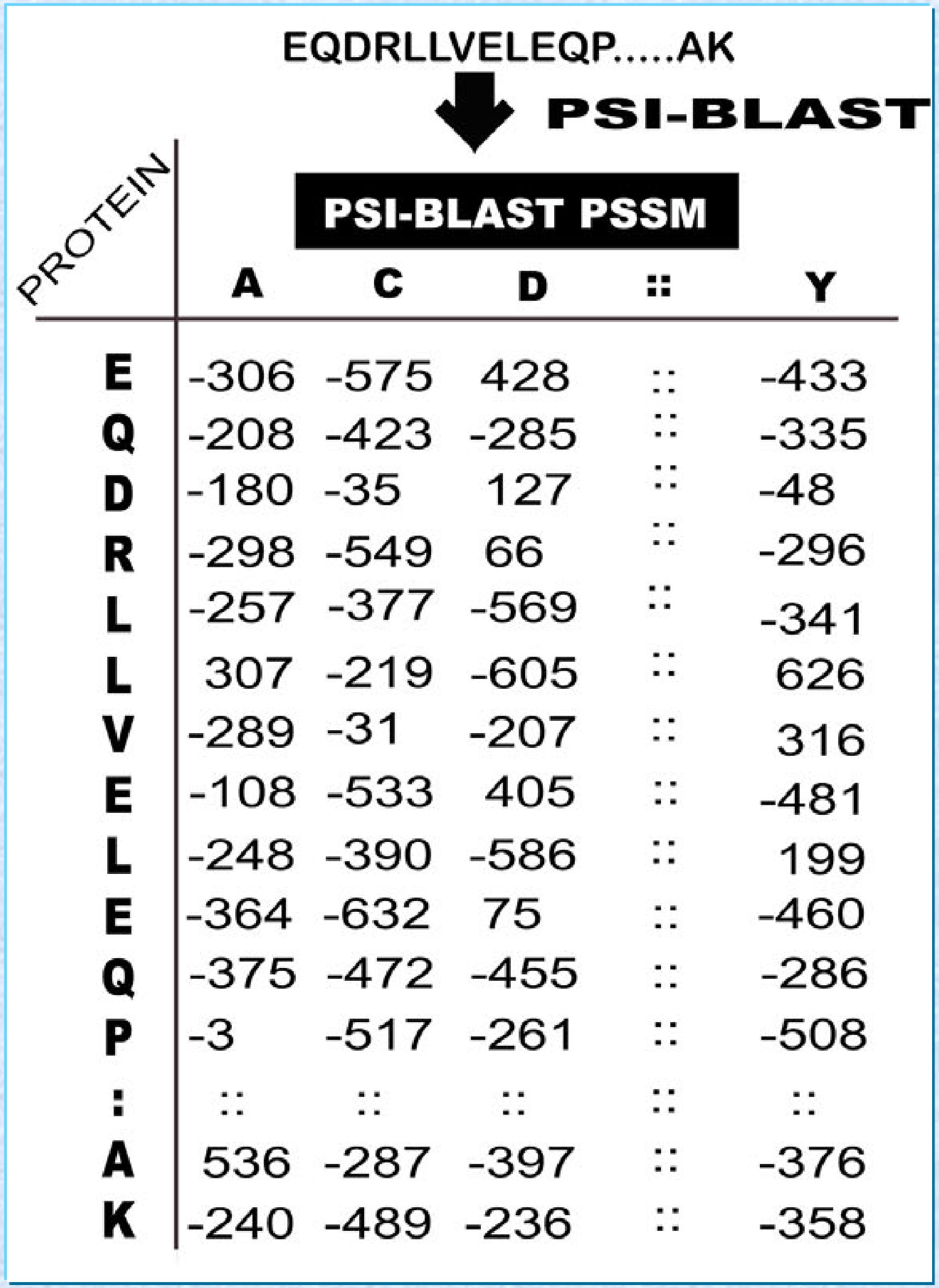
b) Weight Matrix

	1	2	3	4	5	6
A	$\boxed{1.2}$	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	$\boxed{0}$
G	-1.6	$\boxed{.96}$	$\boxed{.96}$	-1.6	$\boxed{.59}$	0
T	-1.6	-1.6	0	$\boxed{.59}$	0	0

test sequence: A G G T G C

Fig. 1. Examples of the simple matrix model for summarizing a DNA alignment. (**a**) An alignment matrix describing the alignment of the four 6-mers on top. The matrix contains the number of times, $n_{i,j}$, that letter i is observed at position j of this alignment. Below the matrix is the consensus sequence corresponding to the alignment (N indicates that there is no nucleotide preference). (**b**) A weight matrix derived from the alignment in (a). The formula used for transforming the alignment matrix to a weight matrix is shown above the arrow. In this formula, N is the total number of sequences (four in this example), p_i is the *a priori* probability of letter i (0.25 for all the bases in this example) and $f_{i,j} = n_{i,j}/N$ is the frequency of letter i at position j. The numbers enclosed in blocks are summed to give the overall score of the test sequence. The overall score is 4.3, which is also the maximum possible score with this weight matrix.

Bioinformatics. 1999;15(7-8):563-77



EQDRLLVELEQP.....AK

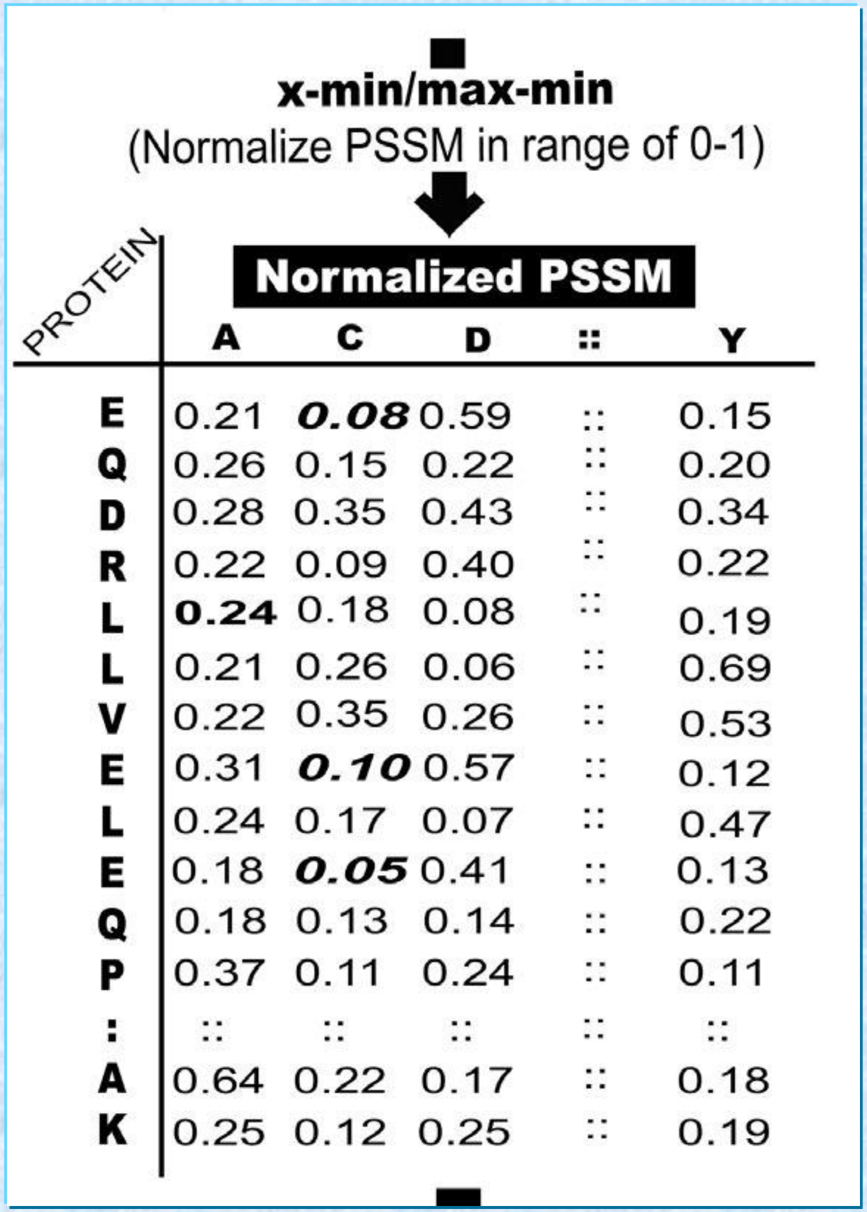
PSI-BLAST

PSI-BLAST PSSM

PROTEIN	A	C	D	::	Y
E	-306	-575	428	::	-433
Q	-208	-423	-285	::	-335
D	-180	-35	127	::	-48
R	-298	-549	66	::	-296
L	-257	-377	-569	::	-341
L	307	-219	-605	::	626
V	-289	-31	-207	::	316
E	-108	-533	405	::	-481
L	-248	-390	-586	::	199
E	-364	-632	75	::	-460
Q	-375	-472	-455	::	-286
P	-3	-517	-261	::	-508
:	::	::	::	::	::
A	536	-287	-397	::	-376
K	-240	-489	-236	::	-358

Method

Given amino acid sequence
get PSI-BLAST results



x-min/max-min

(Normalize PSSM in range of 0-1)

Normalized PSSM

PROTEIN	A	C	D	::	Y
E	0.21	***0.08***	0.59	::	0.15
Q	0.26	0.15	0.22	::	0.20
D	0.28	0.35	0.43	::	0.34
R	0.22	0.09	0.40	::	0.22
L	**0.24**	0.18	0.08	::	0.19
L	0.21	0.26	0.06	::	0.69
V	0.22	0.35	0.26	::	0.53
E	0.31	***0.10***	0.57	::	0.12
L	0.24	0.17	0.07	::	0.47
E	0.18	***0.05***	0.41	::	0.13
Q	0.18	0.13	0.14	::	0.22
P	0.37	0.11	0.24	::	0.11
:	::	::	::	::	::
A	0.64	0.22	0.17	::	0.18
K	0.25	0.12	0.25	::	0.19

Normalize it

X-min/(max-min)

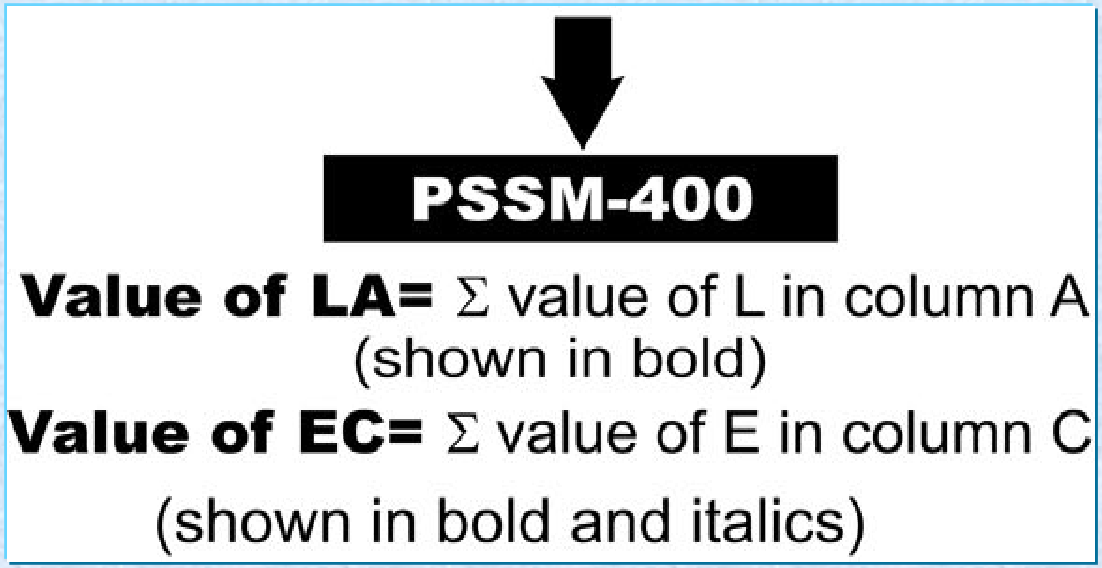
E.g. 5, 6, 9, 2, 1

(5-1)/(9-1) = 4/8 = 0.5

(9-1)/(9-1) = 8/8/ = 1.0

(1-1)/(9-1) = 0/8 = 0.0

PSSM-400



PSSM-400

Value of LA= Σ value of L in column A (shown in bold)

Value of EC= Σ value of E in column C (shown in bold and italics)

EC: 0.08, 0.10, 0.05

LA: 0.24, 0.21, 0.24

QD ?

Applications

Protein secondary structure prediction

Discrimination of proteins belonging to different classes, types etc.

Identifying the binding sites, functionally important residues etc.

Large scale analysis

Non redundant sequences

No two protein sequences have the sequence identity of more than a specific cutoff (say, 40%).

Redundancy cause a bias in any analysis.

E.g. Consider two sequences

ADIKLAAIKL and KILASDPQWE: Average A is 4/20 = 0.20

If one of these sequences appear twice:

ADIKLAAIKL, ADIKLAAIKL and KILASDPQWE: Average A is 7/30= 0.23

(over-represented)

ADIKLAAIKL, KILASDPQWE and KILASDPQWE: Average A is 5/30 = 0.17

(under-represented)

Programs

CD-HIT

Blastclust

PISCES

http://cd-hit.org/

CD-HIT: Cluster Database at High Identity with Tolerance.

The program takes a fasta format sequence database as input and produces a set of 'non-redundant' (nr) representative sequences as output.

It uses clustering algorithm and eliminates the redundant sequences.

The main advantages of this program are given below:

(i) it can handle huge datasets,

(ii) it is easy to download and

(iii) the results can be obtained quickly.

CD-HIT can be used to create the non-redundant dataset of less than 40% sequence identity.

Algorithm

Greedy incremental algorithm: selects representative protein sequence sets

Sequences with the identity of more than the threshold will be discarded.

Longest sequences, the first and proceed with shorter ones.

Sequence identity is the number of identical residues divided by the length of the shorter sequence

Short-word filtering system

Explicit alignment is time consuming

Algorithm without aligning.

Sequences with >90% sequence identity

Decapeptides: query and database (at least 1)

Pentapepides: 85%

Tetrapeptides: 80%

Tripeptides: 75%

Dipeptides: 65% -> efficiency decreases

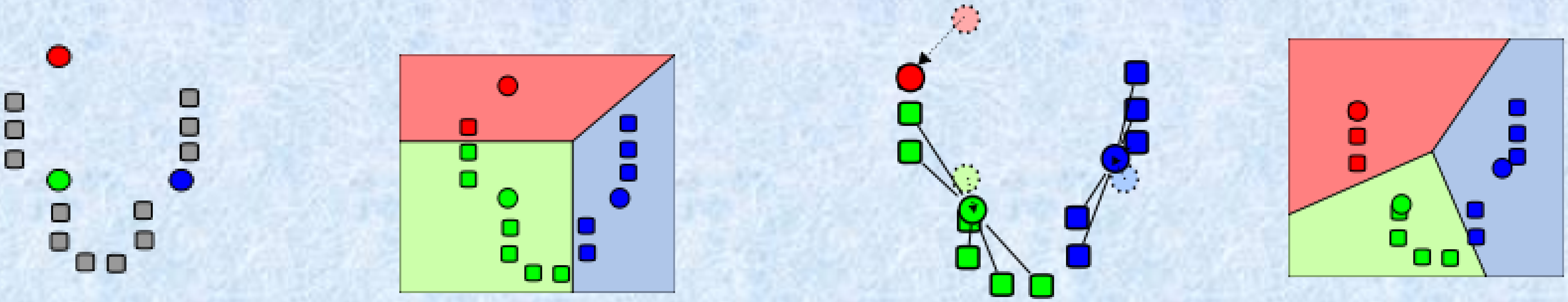
Compare word size and number of same words with sequence identity

Clustering methods

***k*-means clustering** **is a method of cluster analysis which aims to partition** ***n* observations** **into** ***k* clusters** **in which each observation belongs to the cluster with the nearest mean.**

observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), *k*-means clustering aims to partition the *n* observations into *k* sets ($k \leq n$)$\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$



Clustering methods based on composition

Hamming distance

$$D^H = \Sigma\ |Comp(1)_i - comp(2)_i|,\ i=1,20$$

Euclidean distance

$$D^E = \{\Sigma\ [Comp(1)_i - comp(2)_i]^2\}^{1/2}$$

CD-HIT Installation

Installation

Most CD-HIT programs were written in C++.

Download current CD-HIT at http://bioinformatics.org/cd-hit/

Example

cd-hit-v4.5.4-2011-03-07.tgz

Unpack the file with

“tar xvf cd-hit-v4.5.4-2011-03-07.tgz --gunzip”

Change directory by “cd cd-hit-v4.5.4-2011-03-07”

Compile the programs by “make”

Run the program

Run CD-HIT

```
./cd-hit -i db -o db90 -c 0.9 -n 5
```

db: input file name

db90:output file name

0.9, means 90% identity (clustering threshold)

5 is the size of word

Choice of word size:

-n 5 for thresholds 0.7 ~ 1.0

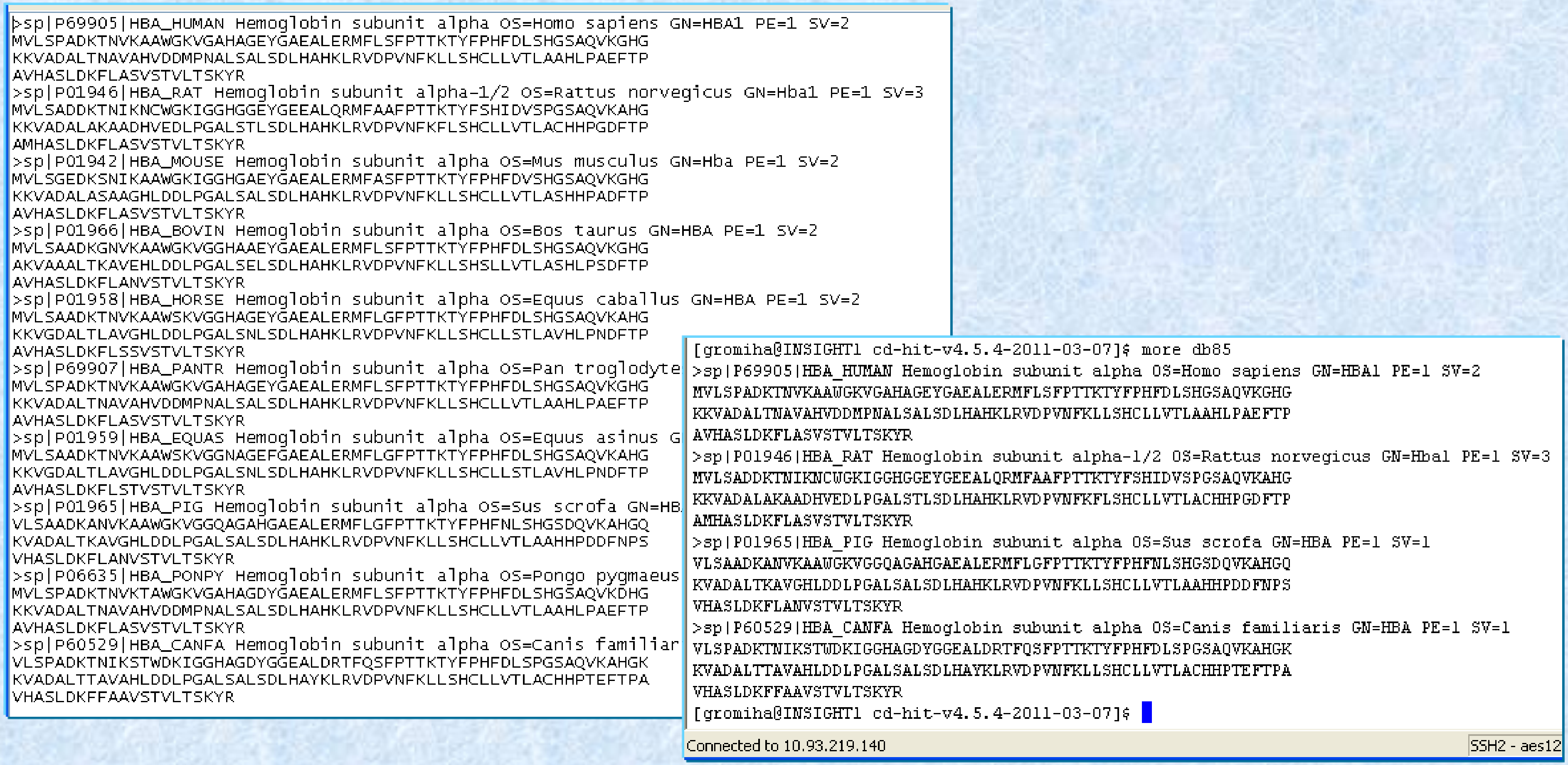
-n 4 for thresholds 0.6 ~ 0.7

-n 3 for thresholds 0.5 ~ 0.6

-n 2 for thresholds 0.4 ~ 0.5

Example

./cd-hit -i hemoglobin_fasta -o db85 -c 0.85 -n 5



```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P01946|HBA_RAT Hemoglobin subunit alpha-1/2 OS=Rattus norvegicus GN=Hba1 PE=1 SV=3
MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG
KKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFKFLSHCLLVTLACHHPGDFTP
AMHASLDKFLASVSTVLTSKYR
>sp|P01942|HBA_MOUSE Hemoglobin subunit alpha OS=Mus musculus GN=Hba PE=1 SV=2
MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P01966|HBA_BOVIN Hemoglobin subunit alpha OS=Bos taurus GN=HBA PE=1 SV=2
MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
AKVAAALTKAVEHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDFTP
AVHASLDKFLANVSTVLTSKYR
>sp|P01958|HBA_HORSE Hemoglobin subunit alpha OS=Equus caballus GN=HBA PE=1 SV=2
MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
KKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTP
AVHASLDKFLSSVSTVLTSKYR
>sp|P69907|HBA_PANTR Hemoglobin subunit alpha OS=Pan troglodyte
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P01959|HBA_EQUAS Hemoglobin subunit alpha OS=Equus asinus G
MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
KKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTP
AVHASLDKFLSTVSTVLTSKYR
>sp|P01965|HBA_PIG Hemoglobin subunit alpha OS=Sus scrofa GN=HB
VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHGQ
KVADALTKAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPDDFNPS
VHASLDKFLANVSTVLTSKYR
>sp|P06635|HBA_PONPY Hemoglobin subunit alpha OS=Pongo pygmaeus
MVLSPADKTNVKTAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P60529|HBA_CANFA Hemoglobin subunit alpha OS=Canis familiar
VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPTEFTPA
VHASLDKFFAAVSTVLTSKYR
```

```
[gromiha@INSIGHT1 cd-hit-v4.5.4-2011-03-07]$ more db85
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P01946|HBA_RAT Hemoglobin subunit alpha-1/2 OS=Rattus norvegicus GN=Hba1 PE=1 SV=3
MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG
KKVADALAKAADHVEDLPGALSTLSDLHAHKLRVDPVNFKFLSHCLLVTLACHHPGDFTP
AMHASLDKFLASVSTVLTSKYR
>sp|P01965|HBA_PIG Hemoglobin subunit alpha OS=Sus scrofa GN=HBA PE=1 SV=1
VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHGQ
KVADALTKAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPDDFNPS
VHASLDKFLANVSTVLTSKYR
>sp|P60529|HBA_CANFA Hemoglobin subunit alpha OS=Canis familiaris GN=HBA PE=1 SV=1
VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTAVAHLDDLPGALSALSDLHAYKLRVDPVNFKLLSHCLLVTLACHHPTEFTPA
VHASLDKFFAAVSTVLTSKYR
[gromiha@INSIGHT1 cd-hit-v4.5.4-2011-03-07]$
```

Blastclust

Blastclust is a program within the standalone BLAST package used to cluster either protein or nucleotide sequences.

The program begins with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster.

In the case of proteins, the blastp algorithm is used to compute the pairwise matches.

The general command to create a set of non-redundant set of protein sequences is *blastclust -i infile -o outfile -p T -L .9 -b T -S 95*,

where infile and outfile are input and output files, respectively.

T stands for protein;

the coverage of the length and sequence identity cutoff are 90% (-L .9) and 95% (-S 95), respectively.

PISCES

PISCES is a protein sequence culling server to produce subsets of non-redundant sequences using Protein Data Bank entries or Uniprot sequences in FASTA format.

Sequence identities for PDB sequences are determined by the combination of Combinatorial Extension structural alignment and PSI-BLAST alignment.

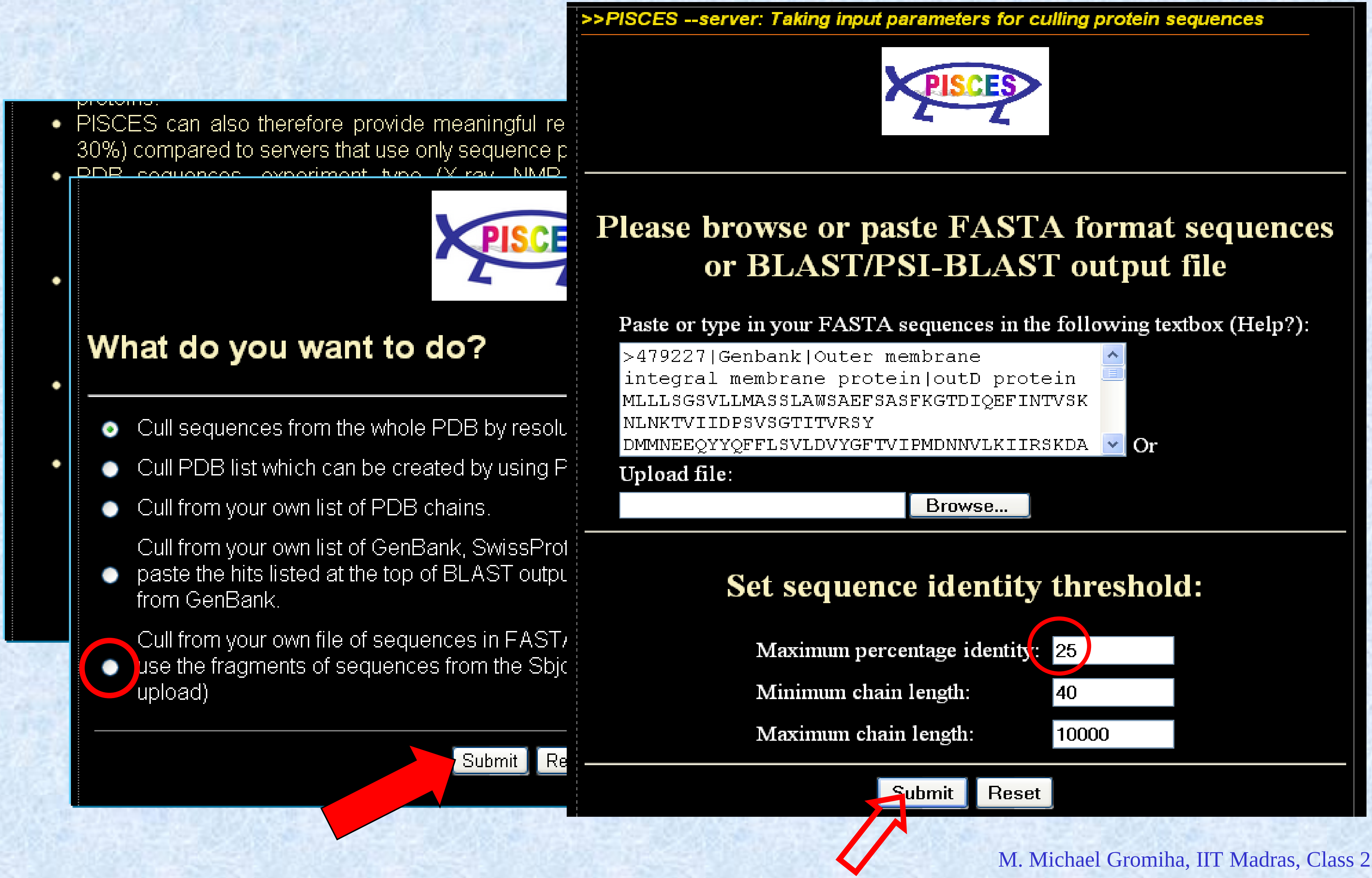
non-PDB sequences are culled with sequence identities from PSI-BLAST. PISCES does not search the non-redundant sequence database, but rather use the user's input sequences as the database.

This server will usually be used to cull a related set of sequences, for instance those from a PSI-BLAST search.

It takes the amino acid sequence in FASTA format and sends the list of non-redundant protein sequences by e-mail.

http://dunbrack.fccc.edu/pisces/

PISCES



proteins.

- PISCES can also therefore provide meaningful re 30%) compared to servers that use only sequence p
- PDB sequences, experiment type (X-ray, NMR

What do you want to do?

◉ Cull sequences from the whole PDB by resolu

○ Cull PDB list which can be created by using P

○ Cull from your own list of PDB chains.

○ Cull from your own list of GenBank, SwissProt paste the hits listed at the top of BLAST outpu from GenBank.

○ Cull from your own file of sequences in FAST/ use the fragments of sequences from the Sbjc upload)

>>PISCES --server: Taking input parameters for culling protein sequences

Please browse or paste FASTA format sequences or BLAST/PSI-BLAST output file

Paste or type in your FASTA sequences in the following textbox (Help?):

```
>479227|Genbank|Outer membrane
integral membrane protein|outD protein
MLLLSGSVLLMASSLAWSAEFSASFKGTDIQEFINTVSK
NLNKTVIIDPSVSGTITVRSY
DMMNEEQYYQFFLSVLDVYGFTVIPMDNNVLKIIRSKDA
```

Or

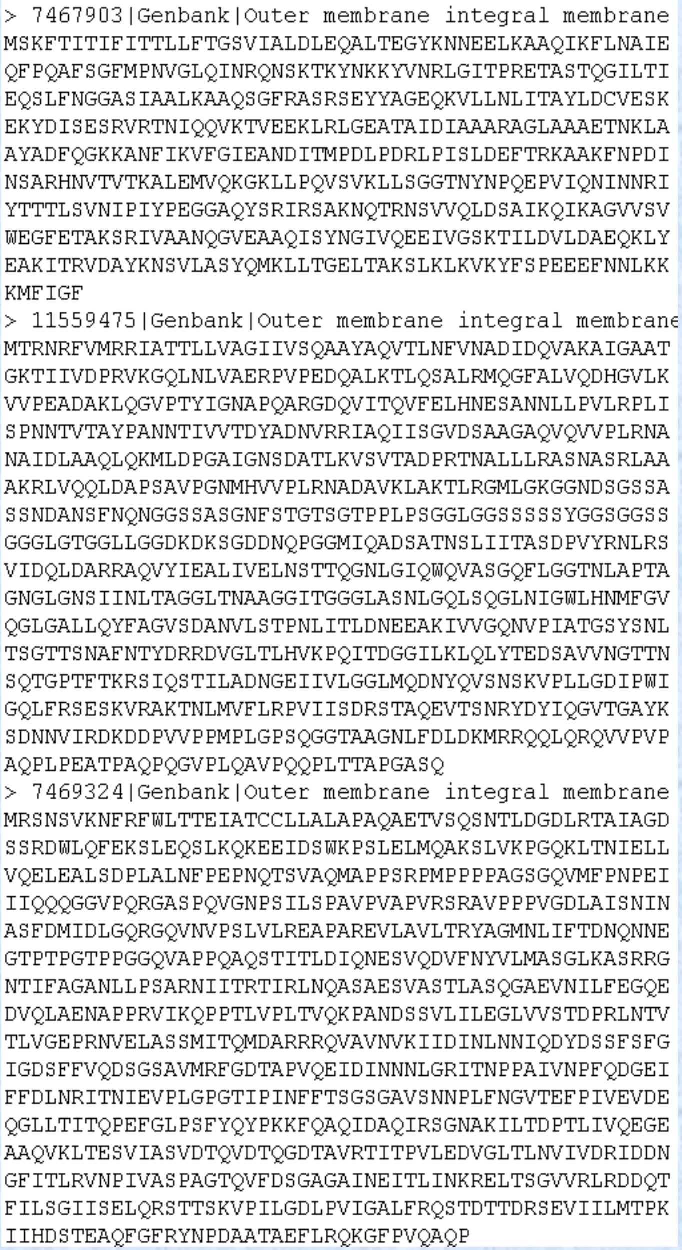
Upload file:

Set sequence identity threshold:

Maximum percentage identity: 25

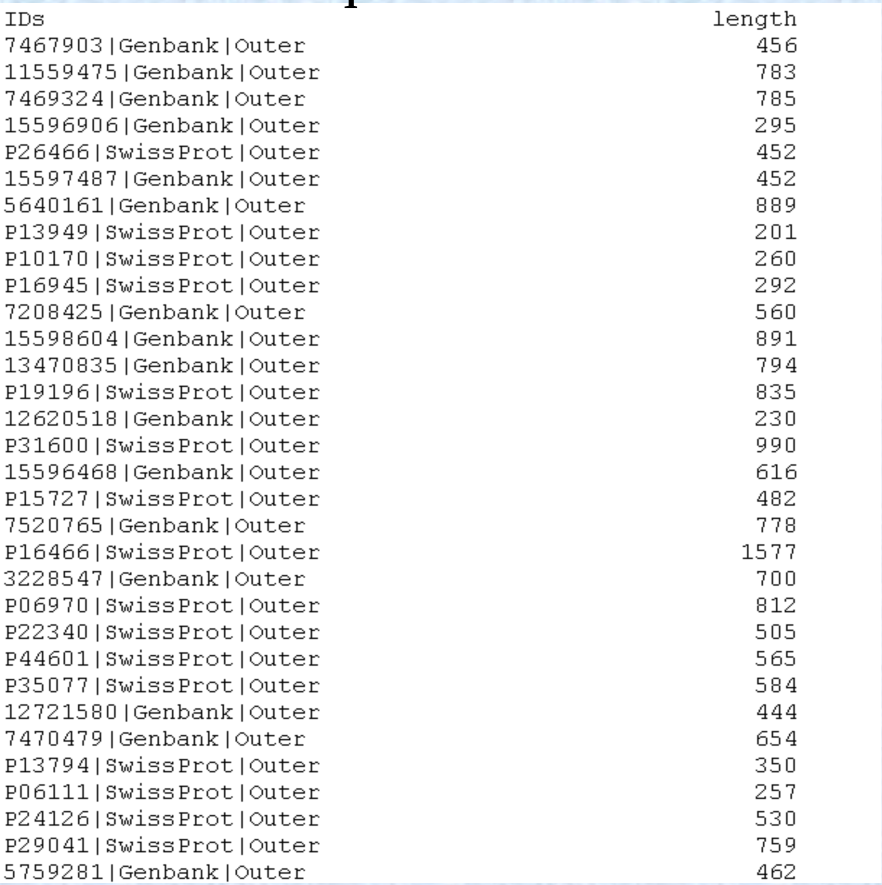
Minimum chain length: 40

Maximum chain length: 10000



```
> 7467903|Genbank|Outer membrane integral membrane
MSKFTITIFITTLLFTGSVIALDLEQALTEGYKNNEELKAAQIKFLNAIE
QFPQAFSGFMPNVGLQINRQNSKTKYNKKYVNRLGITPRETASTQGILTI
EQSLFNGGASIAALKAAQSGFRASRSEYYAGEQKVLLNLITAYLDCVESK
EKYDISESRVRTNIQQVKTVEEKLRLGEATAIDIAAARAGLAAAETNKLA
AYADFQGKKANFIKVFGIEANDITMPDLPDRLPISLDEFTRKAAKFNPDI
NSARHNVTVTKALEMVQKGKLLPQVSVKLLSGGTNYNPQEPVIQNINNRI
YTTTLSVNIPIYPEGGAQYSRIRSAKNQTRNSVVQLDSAIKQIKAGVVSV
WEGFETAKSRIVAANQGVEAAQISYNGIVQEEIVGSKTILDVLDAEQKLY
EAKITRVDAYKNSVLASYQMKLLTGELTAKSLKLKVKYFSPEEEFNNLKK
KMFIGF
> 11559475|Genbank|Outer membrane integral membrane
MTRNRFVMRRIATTLLVAGIIVSQAAYAQVTLNFVNADIDQVAKAIGAAT
GKTIIVDPRVKGQLNLVAERPVPEDQALKTLQSALRMQGFALVQDHGVLK
VVPEADAKLQGVPTYIGNAPQARGDQVITQVFELHNESANNLLPVLRPLI
SPNNTVTAYPANNTIVVTDYADNVRRIAQIISGVDSAAGAQVQVVPLRNA
NAIDLAAQLQKMLDPGAIGNSDATLKVSVTADPRTNALLLRASNASRLAA
AKRLVQQLDAPSAVPGNMHVVPLRNADAVKLAKTLRGMLGKGGNDSGSSA
SSNDANSFNQNGGSSASGNFSTGTSGTPPLPSGGLGGSSSSSYGGSGGSS
GGGLGTGGLLGGDKDKSGDDNQPGGMIQADSATNSLIITASDPVYRNLRS
VIDQLDARRAQVYIEALIVELNSTTQGNLGIQWQVASGQFLGGTNLAPTA
GNGLGNSIINLTAGGLTNAAGGITGGGLASNLGQLSQGLNIGWLHNMFGV
QGLGALLQYFAGVSDANVLSTPNLITLDNEEAKIVVGQNVPIATGSYSNL
TSGTTSNAFNTYDRRDVGLTLHVKPQITDGGILKLQLYTEDSAVVNGTTN
SQTGPTFTKRSIQSTILADNGEIIVLGGLMQDNYQVSNSKVPLLGDIPWI
GQLFRSESKVRAKTNLMVFLRPVIISDRSTAQEVTSNRYDYIQGVTGAYK
SDNNVIRDKDDPVVPPMPLGPSQGGTAAGNLFDLDKMRRQQLQRQVVPVP
AQPLPEATPAQPQGVPLQAVPQQPLTTAPGASQ
> 7469324|Genbank|Outer membrane integral membrane
MRSNSVKNFRFWLTTEIATCCLLALAPAQAETVSQSNTLDGDLRTAIAGD
SSRDWLQFEKSLEQSLKQKEEIDSWKPSLELMQAKSLVKPGQKLTNIELL
VQELEALSDPLALNFPEPNQTSVAQMAPPSRPMPPPPAGSGQVMFPNPEI
IIQQQGGVPQRGASPQVGNPSILSPAVPVAPVRSRAVPPPVGDLAISNIN
ASFDMIDLGQRGQVNVPSLVLREAPAREVLAVLTRYAGMNLIFTDNQNNE
GTPTPGTPPGGQVAPPQAQSTITLDIQNESVQDVFNYVLMASGLKASRRG
NTIFAGANLLPSARNIITRTIRLNQASAESVASTLASQGAEVNILFEGQE
DVQLAENAPPRVIKQPPTLVPLTVQKPANDSSVLILEGLVVSTDPRLNTV
TLVGEPRNVELASSMITQMDARRRQVAVNVKIIDINLNNIQDYDSSFSFG
IGDSFFVQDSGSAVMRFGDTAPVQEIDINNNLGRITNPPAIVNPFQDGEI
FFDLNRITNIEVPLGPGTIPINFFTSGSGAVSNNPLFNGVTEFPIVEVDE
QGLLTITQPEFGLPSFYQYPKKFQAQIDAQIRSGNAKILTDPTLIVQEGE
AAQVKLTESVIASVDTQVDTQGDTAVRTITPVLEDVGLTLNVIVDRIDDN
GFITLRVNPIVASPAGTQVFDSGAGAINEITLINKRELTSGVVRLRDDQT
FILSGIISELQRSTTSKVPILGDLPVIGALFRQSTDTTDRSEVIILMTPK
IIHDSTEAQFGFRYNPDAATAEFLRQKGFPVQAQP
```

Sequence id



IDs	length
7467903|Genbank|Outer	456
11559475|Genbank|Outer	783
7469324|Genbank|Outer	785
15596906|Genbank|Outer	295
P26466|SwissProt|Outer	452
15597487|Genbank|Outer	452
5640161|Genbank|Outer	889
P13949|SwissProt|Outer	201
P10170|SwissProt|Outer	260
P16945|SwissProt|Outer	292
7208425|Genbank|Outer	560
15598604|Genbank|Outer	891
13470835|Genbank|Outer	794
P19196|SwissProt|Outer	835
12620518|Genbank|Outer	230
P31600|SwissProt|Outer	990
15596468|Genbank|Outer	616
P15727|SwissProt|Outer	482
7520765|Genbank|Outer	778
P16466|SwissProt|Outer	1577
3228547|Genbank|Outer	700
P06970|SwissProt|Outer	812
P22340|SwissProt|Outer	505
P44601|SwissProt|Outer	565
P35077|SwissProt|Outer	584
12721580|Genbank|Outer	444
7470479|Genbank|Outer	654
P13794|SwissProt|Outer	350
P06111|SwissProt|Outer	257
P24126|SwissProt|Outer	530
P29041|SwissProt|Outer	759
5759281|Genbank|Outer	462