Name: Saurabh Khandagale

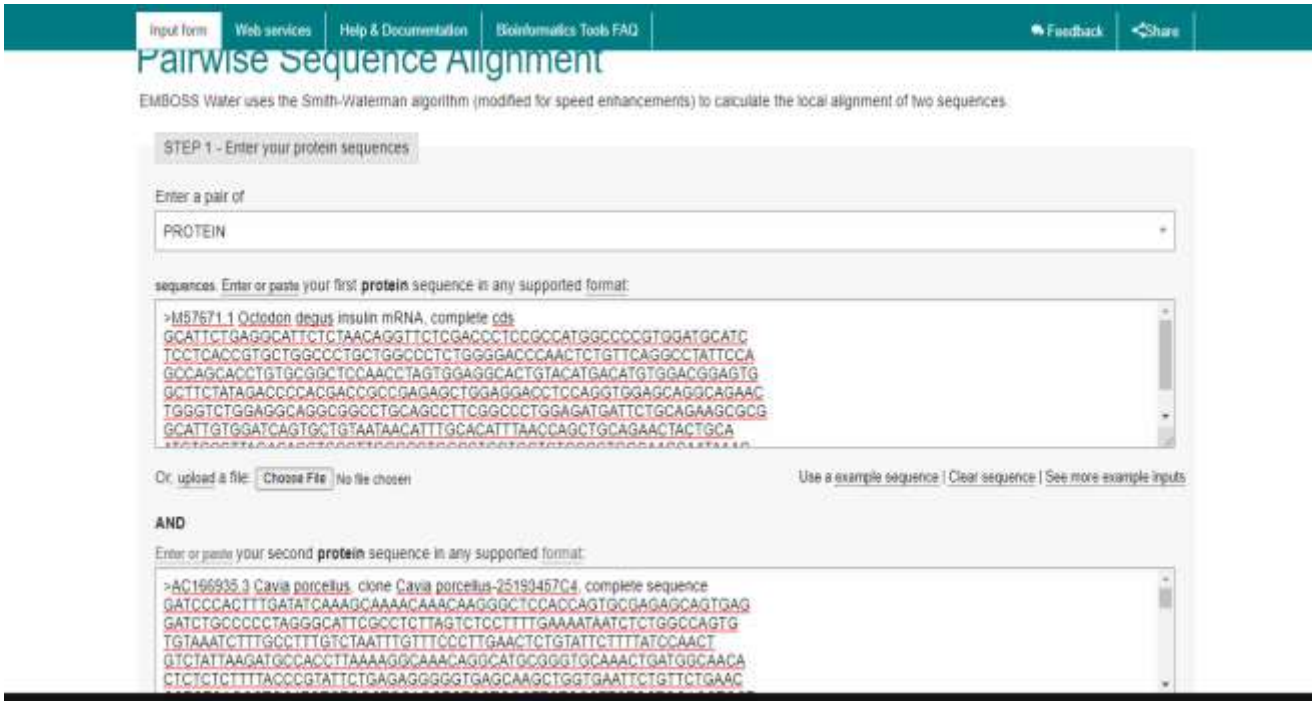
Roll number/batch : 46

Year : 4th Year CSE Shift 2

Subject : Bio-Informatics

Semester: VII

1) Take the Insulin protein (amino acid) sequences of two organisms like Octodon degus (AAA40590.1) which is a species of rodent (commonly called rat) and Cavia Porcellus (AAA37041.1), commonly called Guinea pig, also a species of rodent rat. Align both of the protein sequences using local alignment method based on a online tool "EMBOSS-water"(http://www.ebi.ac.uk/Tools/psa/emboss_water/) and find out the similarity percentage between the two protein sequences.

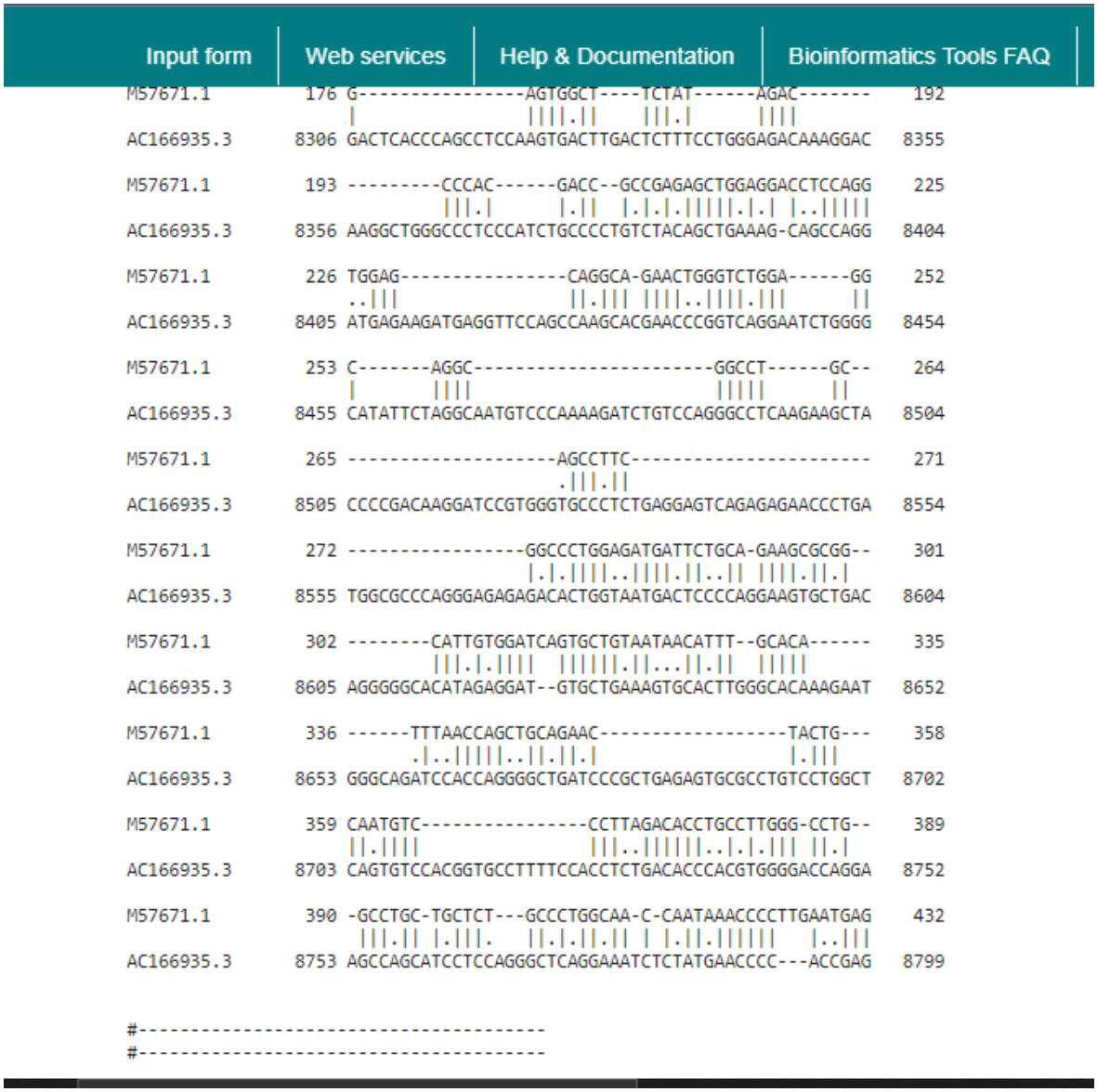


Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ

```
########################################
# Program: water
# Rundate: Mon 18 Oct 2021 07:16:59
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20211018-071656-0609-21294862-p2m.asequence
#    -bsequence emboss_water-I20211018-071656-0609-21294862-p2m.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: M57671.1
# 2: AC166935.3
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 900
# Identity:     318/900 (35.3%)
# Similarity:   318/900 (35.3%)
# Gaps:         474/900 (52.7%)
# Score: 1338.0
#
#
#=======================================

M57671.1           1 GC-----ATTCTGAGGCATTCTCTAAC-------------AGGTTCTCGA     32
                       ||     .|||||.|..||.|.||.||             .|||.|||..
AC166935.3      7906 GCCCCCACTTCTGGGTTATGCCCTGACTCAGACCCACCCACGGTCCTCTC   7955
```



⇨ SIMILARITY PERCENTAGE is found to be – 35.3%

2) Given a protein sequence, find out which protein it is and from which organism it belongs using BLAST Program.

Protein Sequence

MERGVRRGAALVAAWRSLWERGGLALFRPQCRTGCGACRVQGTRPFSLSAAAS
AVLG
LGSWGGDSGKQKLTLQDVAELIRKKECRRVVVMAGAGIS
TPSGIPDFRSPGSGLYSNLEQYNIPYPEAIFELAYFFINPKPFFTLAKELYPGNYRPN
YAHY
FLRLLHDKGLLLRLYTQNIDGLERVAGIPPDRLVEAHGT
FATATCTVCRRKFPGEDFRGDVMADKVPHCRVCTGIVKPDIVFFGEELPQRFFLH
MTDF
PMADLLFVIGTSLEVEPFASLAGAVRNSVPRVLINRDLV
GPFAWQQRYNDIAQLGDVVTGVEKMVELLDWNEEMQTLIQKEKEKLDAKDK

⇨ ORGANISM : Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates;
Haplorrhini;
Catarrhini; Hominidae; Homo.

From the given details, find

- What is the length of the query sequence ?

⇨ Length of Query = 346

- What is the number of sequences in the database searched ?

⇨

- What is the E-value ?

Best hits

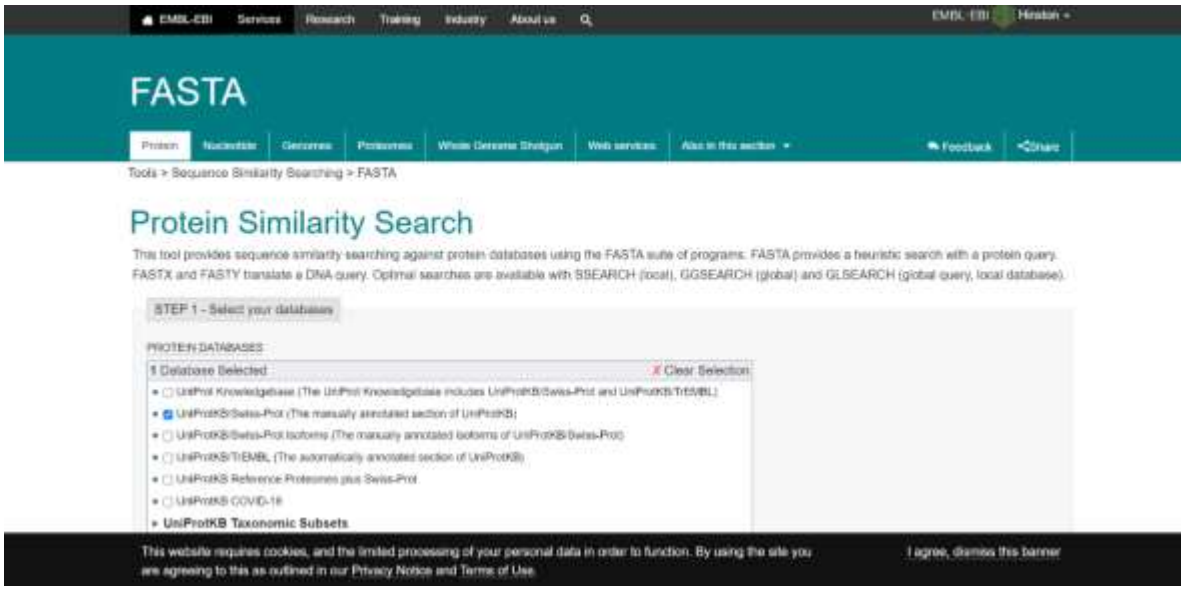
Select All None Selected:0

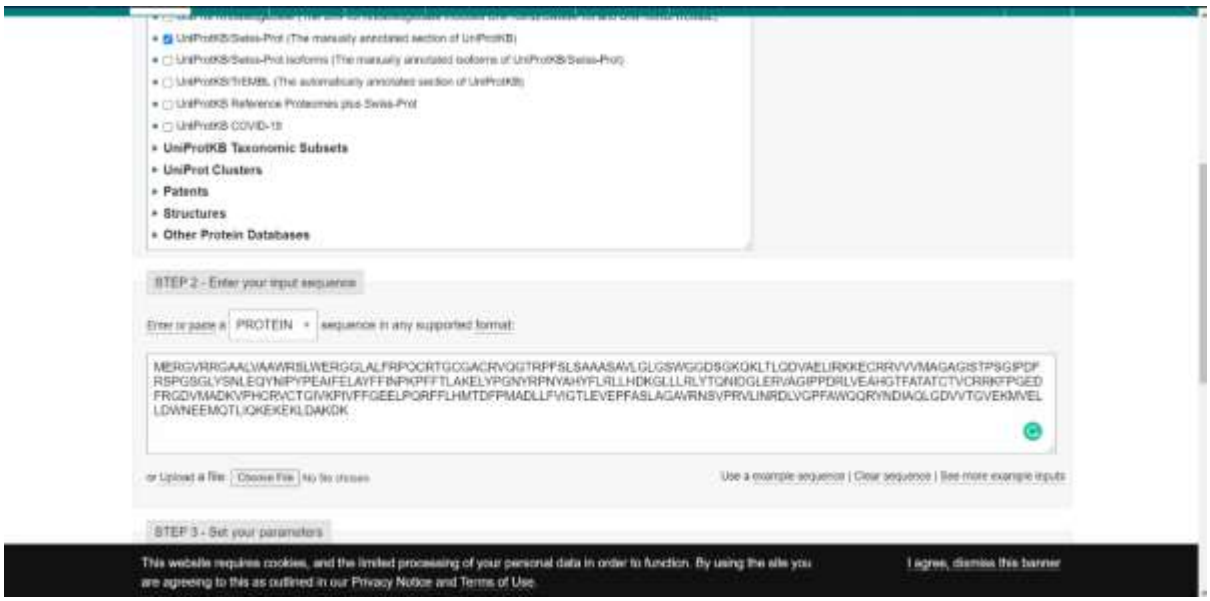
Alignments

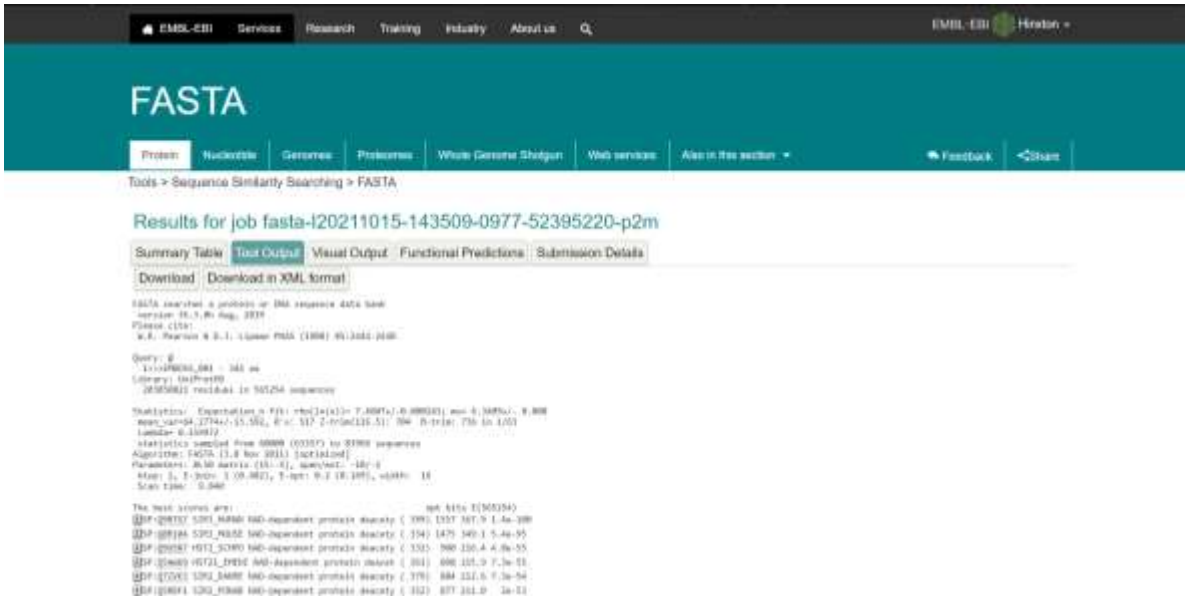
	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	NAD-dependent protein deacetylase sirtuin-3, mitochondrial isoform a [Homo sapiens]	468	468	80%	1e-164	78.93%	NP_036371.1
☐	NAD-dependent protein deacetylase sirtuin-3 isoform 3 [Mus musculus]	444	444	80%	3e-156	73.93%	NP_001171275.1
☐	NAD-dependent protein deacetylase sirtuin-3, mitochondrial isoform X1 [Danio rerio]	419	419	79%	6e-146	71.27%	XP_009301762.1
☐	NAD-dependent protein deacetylase sirtuin-3, mitochondrial isoform f precursor [Homo sapiens]	457	457	79%	1e-161	78.83%	NP_001357244.1
☐	NAD-dependent protein deacetylase sirtuin-3, mitochondrial isoform c [Homo sapiens]	460	460	80%	3e-161	77.50%	NP_001357239.1

Additional BLAST Hits

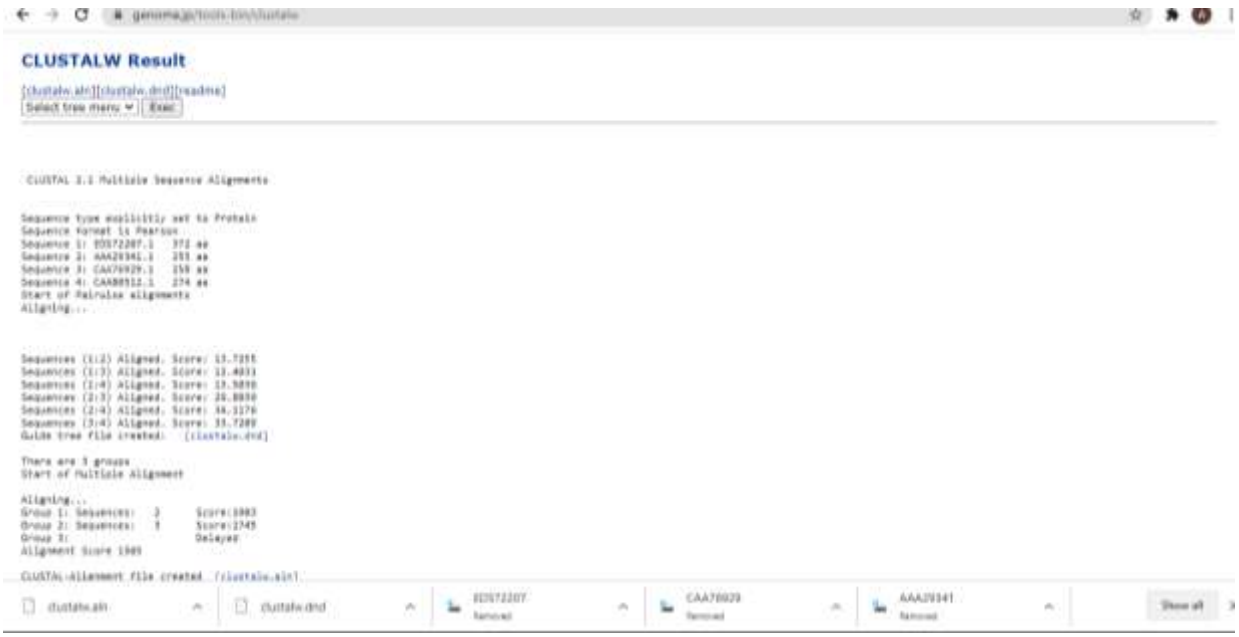
3) Given a protein sequence find out which protein it is and from which organism it belongs from FASTA program based on the identity.Given a protein sequence find out which protein it is and from which organism it belongs from FASTA program based on the identity. Also find the E-value for the given sequence.







4) Collect the protein sequences CAA80512, AAA29341, CAA76929, and EDS72207 from the enterz database, keep the sequences together in FASTA format file, align the sequences each other and report the pair wise score using CLUSTALW?


CLUSTALW Result
(clustalw.aln)[clustalw.dnd][readme]
Select tree menu
Exec
CLUSTAL 2.1 Multiple Sequence Alignments
Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: EDS72207.1 372 aa
Sequence 2: AAA29341.1 353 aa
Sequence 3: CAA76929.1 159 aa
Sequence 4: CAA88512.1 274 aa
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score: 13.7255
Sequences (1:3) Aligned. Score: 13.4831
Sequences (1:4) Aligned. Score: 13.5859
Sequences (2:3) Aligned. Score: 20.8850
Sequences (2:4) Aligned. Score: 46.1176
Sequences (3:4) Aligned. Score: 33.7289
Guide tree file created: (clustalw.dnd)
There are 3 groups
Start of Multiple Alignment
Aligning...
Group 1: Sequences: 2 Score:1983
Group 2: Sequences: 3 Score:2745
Group 3: Delayed
Alignment Score 1985
CLUSTAL-Alignment file created (clustalw.aln)
clustalw.aln
clustalw.dnd
EDS72207
CAA76929
AAA29341
Show all

CLUSTAL-Alignment file created (clustalw.aln)
clustalw.aln
CLUSTAL 2.1 multiple sequence alignment
clustalw.aln
clustalw.dnd
EDS72207
CAA76929
AAA29341
Show all

clustalw.dnd
Select tree menu
Exec
clustalw.aln
clustalw.dnd
Show all