BT3040 – BIOINFORMATICS – Assignment 3

Submitted by Sahana (BE17B038)

Question1 –

AA sequence –

```
>sp|P21796|VDAC1_HUMAN Voltage-dependent anion-selective channel
protein 1 OS=Homo sapiens OX=9606 GN=VDAC1 PE=1 SV=2

MAVPPTYADLGKSARDVFTKGYGFGLIKLDLKTKSENGLEFTSSGSANTETTKVTGSLET
KYRWTEYGLTFTEKWNTDNTLGTEITVEDQLARGLKLTFDSSFSPNTGKKNAKIKTGYKR
EHINLGCDMDFDIAGPSIRGALVLGYEGWLAGYQMNFETAKSRVTQSNFAVGYKTDEFQL
HTNVNDGTEFGGSIYQKVNKKLETAVNLAWTAGNSNTRFGIAAKYQIDPDACFSAKVNNS
SLIGLGYTQTLKPGIKLTLSALLDGKNVNAGGHKLGLGLEFQA.
```

Algorithm –

1. Go to UniProt. Search for "human mitochondrial beta barrel membrane protein VDAC1.
2. https://www.uniprot.org/uniprot/P21796

Function – Forms a channel through the mitochondrial outer membrane and also the plasma membrane. The channel at the outer mitochondrial membrane allows diffusion of small hydrophilic molecules; in the plasma membrane it is involved in cell volume regulation and apoptosis. It adopts an open conformation at low or zero membrane potential and a closed conformation at potentials above 30-40 mV. The open state has a weak anion selectivity whereas the closed state is cation-selective. May participate in the formation of the permeability transition pore complex (PTPC) responsible for the release of mitochondrial products that triggers apoptosis.

There are **19 transmembrane segments** present in this protein.

Question 2 –

Total number of sequences = 91,451

50% identity = 11,118 clusters

90% identity = 21,604 clusters

100% identity = 49,861 clusters

Algorithm –

1. Search for "transcription factors" in UniProt.
   https://www.uniprot.org/uniprot/?query=%22transcription+factors%22&sort=score
2. Select cluster identities for specific values.

Question 3 –

There are 188,436 sequences of "homo sapiens" in UniProt.

Sequence identity	Number of clusters
100%	140,115
90%	84,972
50%	65,485

Question 4 –

Search query = reviewed:yes AND organism:"Mus musculus (Mouse) [10090]"

17,027 sequences are manually annotated for *"Mus Musculus"*.

Search query = database:(type:pdb) AND reviewed:yes AND organism:"Mus musculus (Mouse) [10090]"

There are 1,873 sequences from the above, which also have 3D structure in PDB.

Question 5 –

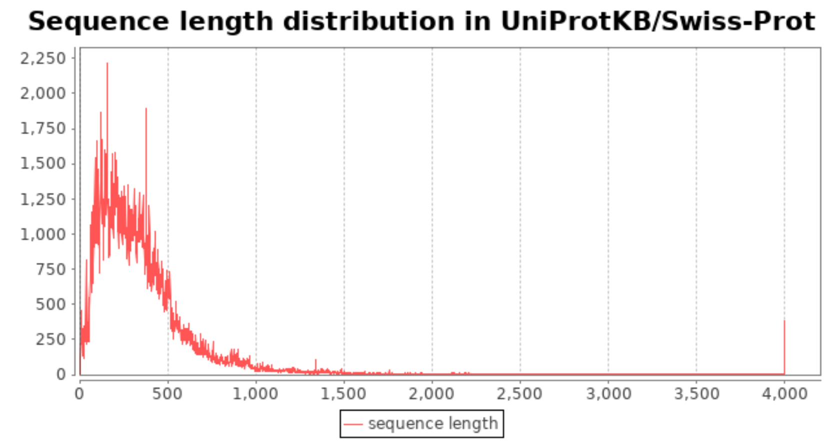
1,770 out of 1,873 identifiers from UniProtKB AC/ID were successfully mapped to 1,770 STRING IDs.

Algorithm –

1. Search in UniProt with search query –
2. database:(type:pdb) AND reviewed:yes AND organism:"Mus musculus (Mouse) [10090]"
3. Select only the entry column and download the identifiers as a list.
4. In Retrieve/ID Mapping, paste these identifiers.
5. Under select options:
    a. From – UniProtKB
    b. To – STRING…. And submit.

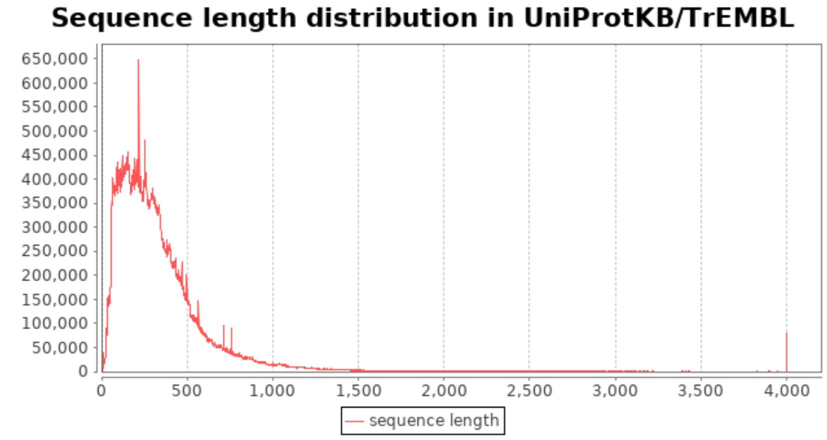
Question 6 – (https://www.uniprot.org/statistics/Swiss-Prot)

(a)



Sequence length distribution in UniProtKB/Swiss-Prot

The shortest sequence is P0DPR3 at 2 AA while the longest sequence is A2ASS6 at 35,213 AA



Sequence length distribution in UniProtKB/TrEMBL

The shortest sequence is A0A1Y7VI41 at 7 AA while the longest sequence is A0A5A9P0L4 at 45,354 AA

Inference – There is a lot of sequences with length between 200-400 Amino Acids. The frequency of sequences with longer length of Amino acids is lesser.

(b) As per Swiss_Prot,

The shortest sequence in UniProtKB = 2 Amino acids, which sequence ID = P0DPR3

The longest sequence in UniProtKB = 35,213 Amino acids with sequence ID = A2ASS6

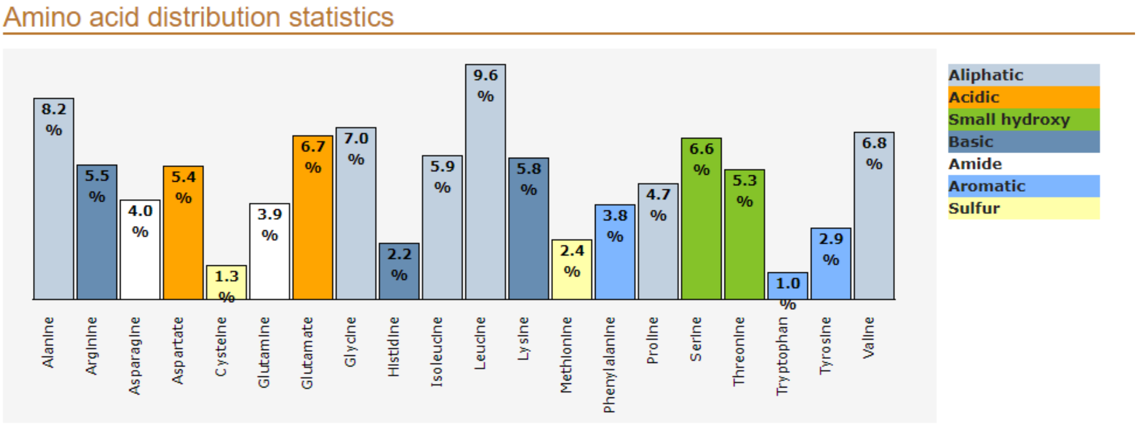
As per TrEMBL,

The shortest sequence in UniProtKB = 7 Amino acids, which sequence ID = A0A1Y7VI41

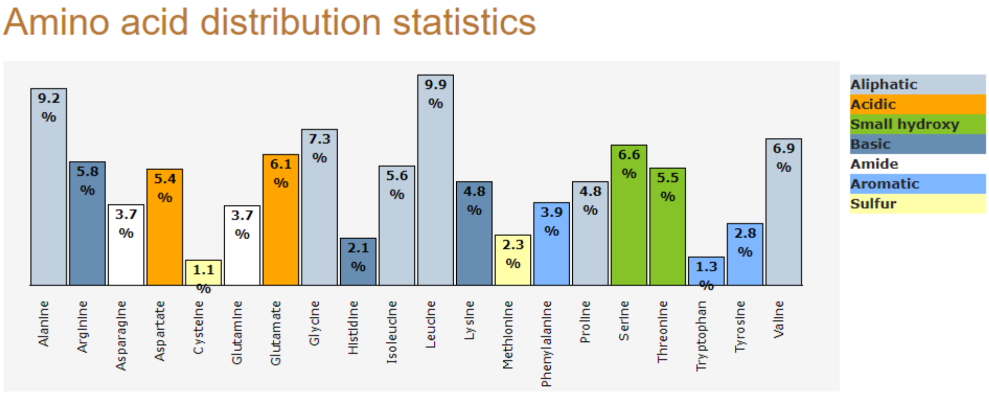
The longest sequence in UniProtKB = 45,354 Amino acids with sequence ID = A0A5A9P0L4

(c)

As per Swiss_Prot



As per TrEMBL,



Question 7 -

Human haemoglobin beta chain – UniProt ID – P68871

'MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH'

Chicken haemoglobin beta chain – UniProt ID – P02112

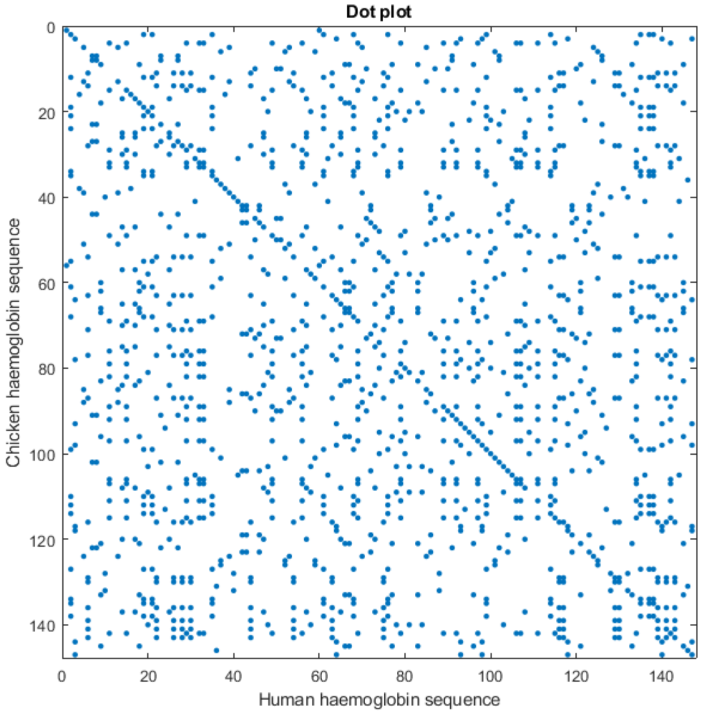
'MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM
VRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS
KDFTPECQAAWQKLVRVVAHALARKYH'

Matlab code –

```
function[] = dot_plot(A,B)
    n = length(A);
    m = length(B);
    D= zeros(n,m);
    for i= 1:n
        for j = 1:m
            if A(i)==B(j)
                D(i,j)=1;
            end
        end
    end
    final = '';
    len = 0;
    l = max(n,m);
    for k = 1:l
        if A(k)==B(k)
            final = append(final,A(k));
            len=len+1;
        else
            final = append(final,'-');
        end
    end
    spy(D)
    title('Dot plot');
    xlabel('Human haemoglobin sequence');
    ylabel('Chicken haemoglobin sequence');
    fprintf('The most common segment between both the sequences = ');
    disp(final)
    fprintf('Length of common segment = ');
    disp(len)
end
```

Output –

```
>> dot_plot(h,c)
The most common segment between both the sequences = MVH-T-EEK---T-LWGKVNV-
E-G-EAL-RLL-VYPWTQRFF-SFG-LS-P-A--GNP-V-AHGKKVL--F-D----LDN-K-TF--
LSELHCDKLHVDPENFRLLG--L--VLA-HF-K-FTP--QAA-QK-V--VA-ALA-KYH
Length of common segment =    102
```


Dot plot
Chicken haemoglobin sequence
Human haemoglobin sequence