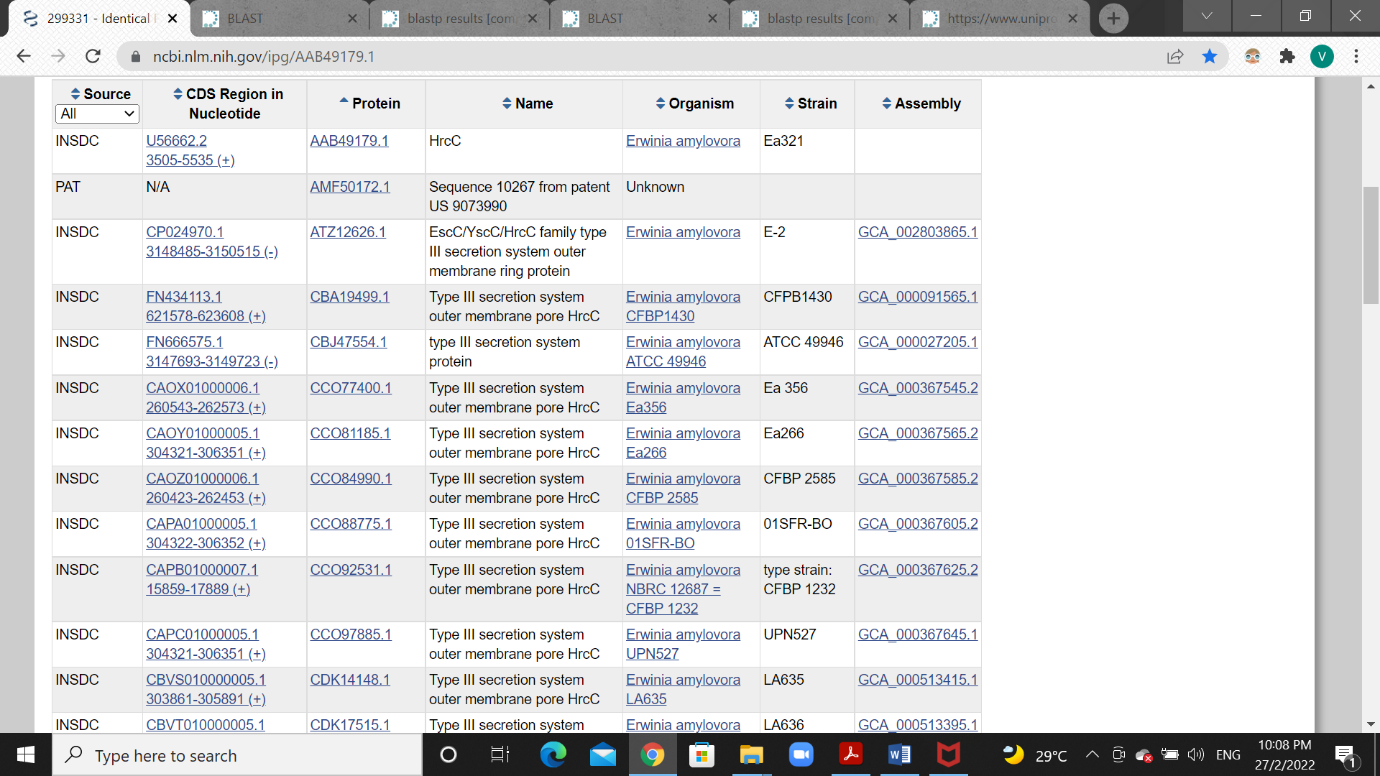
***Practicals – 5***

*-BS19B032*

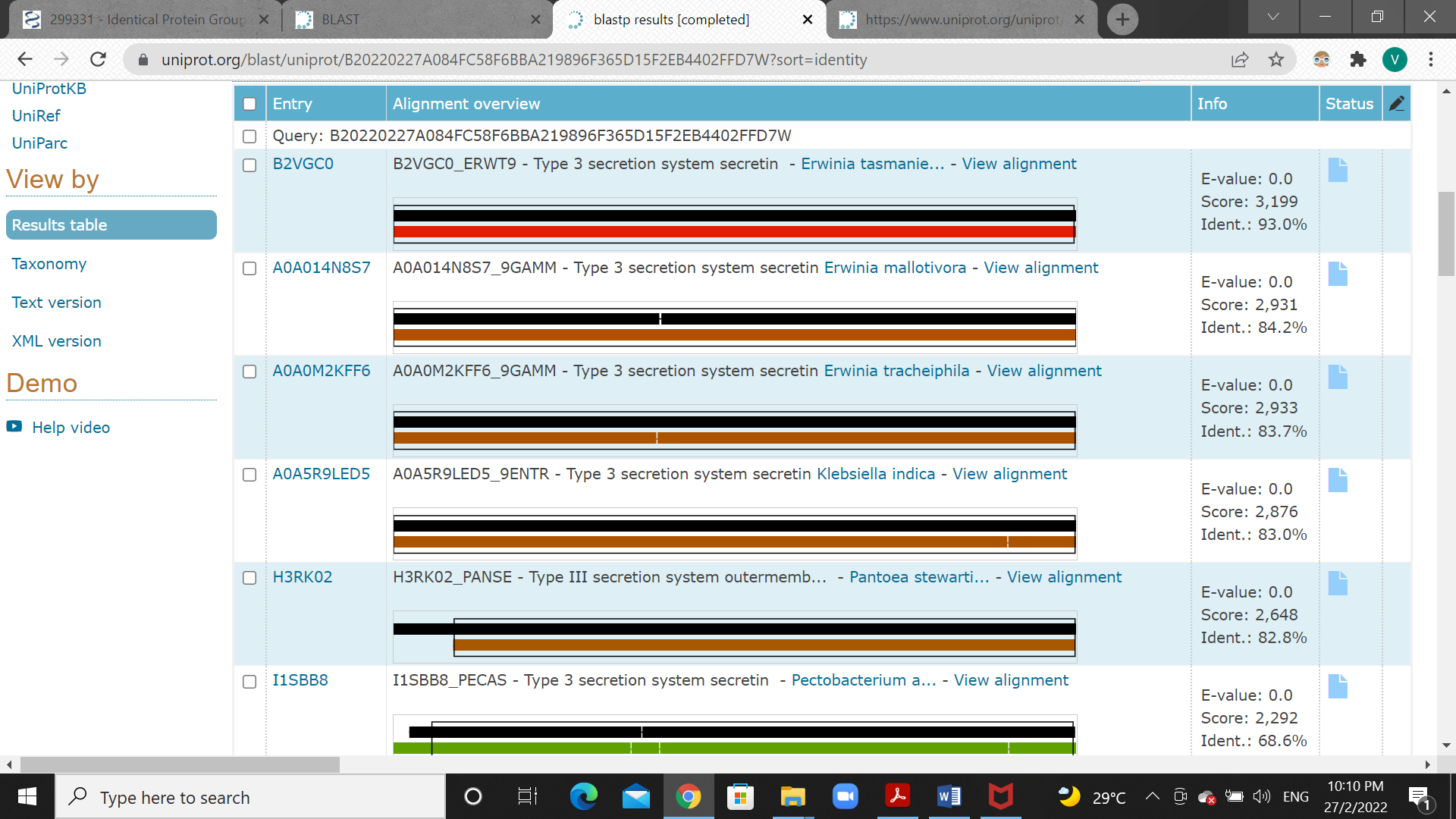
*-R. Vasantha Kumar*

1) For the given sequence, the similar proteins were searched:

**nr:**

****

**SWISS-PROT:**



In nr, a total of 200 similar proteins were found. Whereas, in SWISS-PROT blast 250 sequences were found, it is because the maximum target sequences is 250. But, most of the sequences are with low sequence identity.

2) The algorithm parameters used for searching similar proteins in question 1 are:

* Max target sequences = 250
* Expect Threshold = 0.05
* Word size = 28
* Match score = 1
* Mismatch score = -2
* Gap costs = Linear

3) the sequence identity of the query sequence (given in Q1) with AAK81929.1 is:

>AAK81929.1 RscC [Pseudomonas fluorescens]

MHNKISKHTCLHIDPPDTSSRRAKWQWLVLLGCIMAPAHNLLAAIPAEWKNTAYAYEADHKPLREVLEDF

AQTFGTQLQIEGLLEGDVNGKIRANTPQSMLDRLGVEHRFQWYLYNNTLFVSTLDQQESARLEVSSETIS

DLKQALTDIGLLDSRFGWGELPEDGVVLVSGPKTYIDQIKQFSSKRRSADEKQSVLSFPLKFANAADRKV

DYRGEKLVVPGVANILRGLLEPRSASTLTGMSQPDSSQPSPLTPNVPRLGNPLLGQMLGANGNAGQLDTG

PTVTPRAPVSKSRIRVEADVRNNAVLIYDLPERQAMYRDLITQLDVARKLIEIDAIILDIERTQLREFGV

NWGFQNSRFRGGVNMAPGTSSQVSIDHRDRFYADMPSTGGQGPATMVSNPSVLTLENQPAVIDFNRTQYI

SPGRDYATILPVTVGTSLQVVPRVTTGRGVHQIHLVVDIEDGNLDETNPERDPNHLDVRRGKVSTQAVMQ

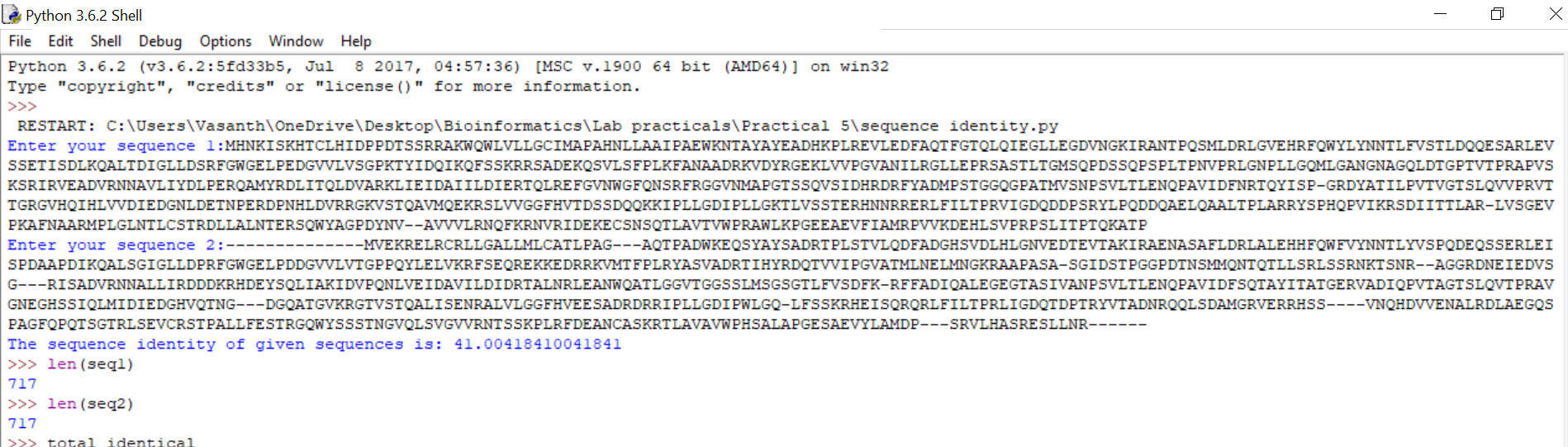
EKRSLVVGGFHVTDSSDQQKKIPLLGDIPLLGKTLVSSTERHNNRRERLFILTPRVIGDQDDPSRYLPQD

DQAELQAALTPLARRYSPHQPVIKRSDIITTLARLVSGEVPKAFNAARMPLGLNTLCSTRDLLALNTERS

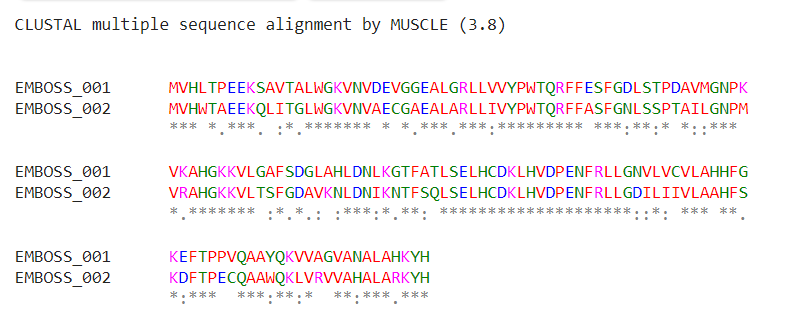
QWYAGPDYNVAVVVLRNQFKRNVRIDEKECSNSQTLAVTVWPRAWLKPGEEAEVFIAMRPVVKDEHLSVP

RPSLITPTQKATP

The sequence identity is 41%.



4) To find how far are hemoglobin sequences in human and chicken similar, I used muscle multiple alignment of sequences:



From above results, we could find that these two sequences are very similar. Out of sequence length of 147, only 23 of them are not similar.

5) The program to list all the matching pentapeptides (which occur in both the sequences) and their frequency of occurrence in given sequences is:

seq1 = input("Enter your sequence1:")

seq2 = input("Enter your sequence2:")

print("The Matching Pentapeptides are:")

count = 0

while(count<=(len(seq1)-5)):

pen\_seq1 = seq1[count:count+5]

if pen\_seq1 in seq2:

print(pen\_seq1)

count2 = 0

freq1 = 1

freq2 = 0

print("The frequency of occurence in sequence 1 is:",freq1)

while(count2<=(len(seq2)-5)):

check\_seq = seq2[count2:count+5]

if(check\_seq == pen\_seq1):

freq2 = freq2 + 1

count2 = count2 + 1

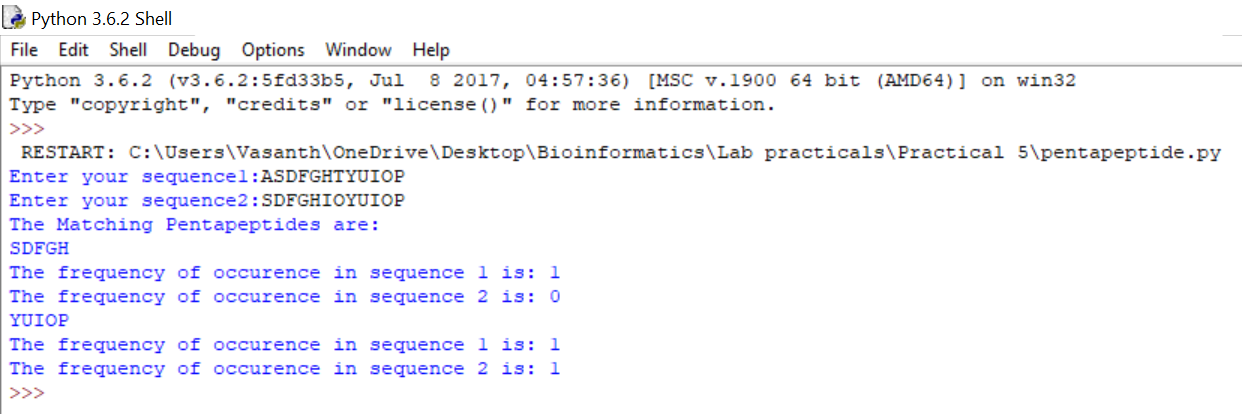
print("The frequency of occurence in sequence 2 is:",freq2)

count = count + 1

For example, the result for these sequences is:

ASDFGHTYUIOP

SDFGHIOYUIOP



6) The program to compute sequence identity, similarity, query coverage and gap percentage from the alignment of human and chicken hemoglobin sequences is:

human\_seq='MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH'

chick\_seq='MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH'

total\_identical = 0

total\_similarity = 0

gap1 = 0

gap2 = 0

for i in range(len(human\_seq)):

if(human\_seq[i]=='-'):

gap1 = gap1 + 1

if(chick\_seq[i]=='-'):

gap2 = gap2 + 1

if(human\_seq[i]==chick\_seq[i]):

total\_identical = total\_identical + 1

elif((human\_seq[i]=='A') or (human\_seq[i]=='I') or (human\_seq[i]=='L') or (human\_seq[i]=='M') or (human\_seq[i]=='F') or (human\_seq[i]=='W') or (human\_seq[i]=='V')):

if((chick\_seq[i]=='A') or (chick\_seq[i]=='I') or (chick\_seq[i]=='L') or (chick\_seq[i]=='M') or (chick\_seq[i]=='F') or (chick\_seq[i]=='W') or (chick\_seq[i]=='V')):

total\_similarity = total\_similarity + 1

elif((human\_seq[i]=='K') or (human\_seq[i]=='R')):

if((human\_seq[i]=='K') or (human\_seq[i]=='R')):

total\_similarity = total\_similarity + 1

elif((human\_seq[i]=='S') or (human\_seq[i]=='T')):

if((human\_seq[i]=='S') or (human\_seq[i]=='T')):

total\_similarity = total\_similarity + 1

elif((human\_seq[i]=='H') or (human\_seq[i]=='Y')):

if((human\_seq[i]=='H') or (human\_seq[i]=='Y')):

total\_similarity = total\_similarity + 1

sequence\_identity = (total\_identical/len(human\_seq))\*100

print("The sequence identity of given sequences is:",sequence\_identity)

similarity = ((total\_identical+total\_similarity)/len(human\_seq))\*100

print("The similarity of given sequences is:",similarity)

gap\_percentage1 = (gap1/len(human\_seq))\*100

gap\_percentage2 = (gap2/len(chick\_seq))\*100

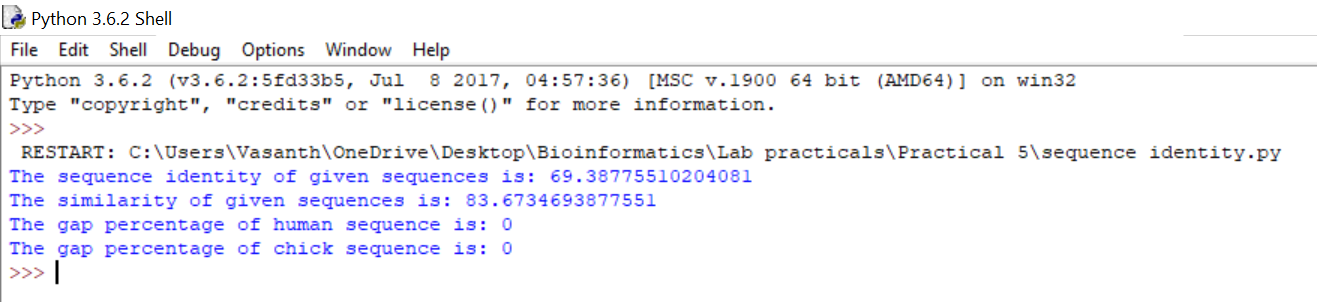
print("The gap percentage of human sequence is:",gap1)

print("The gap percentage of chick sequence is:",gap2)

The sequence identity is 69.38%.

The similarity is 83.67%.

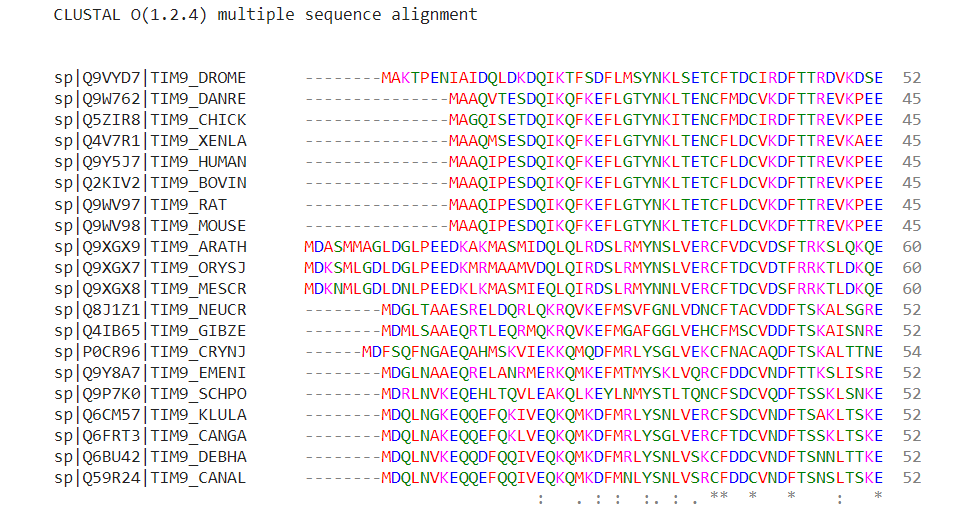
The gap percentage of both the sequences is 0.

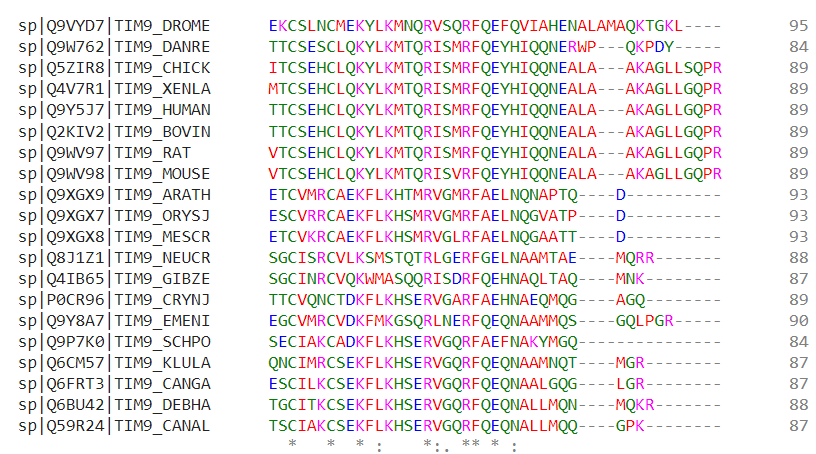


7) 20 TIM barrel proteins from different organisms are taken.

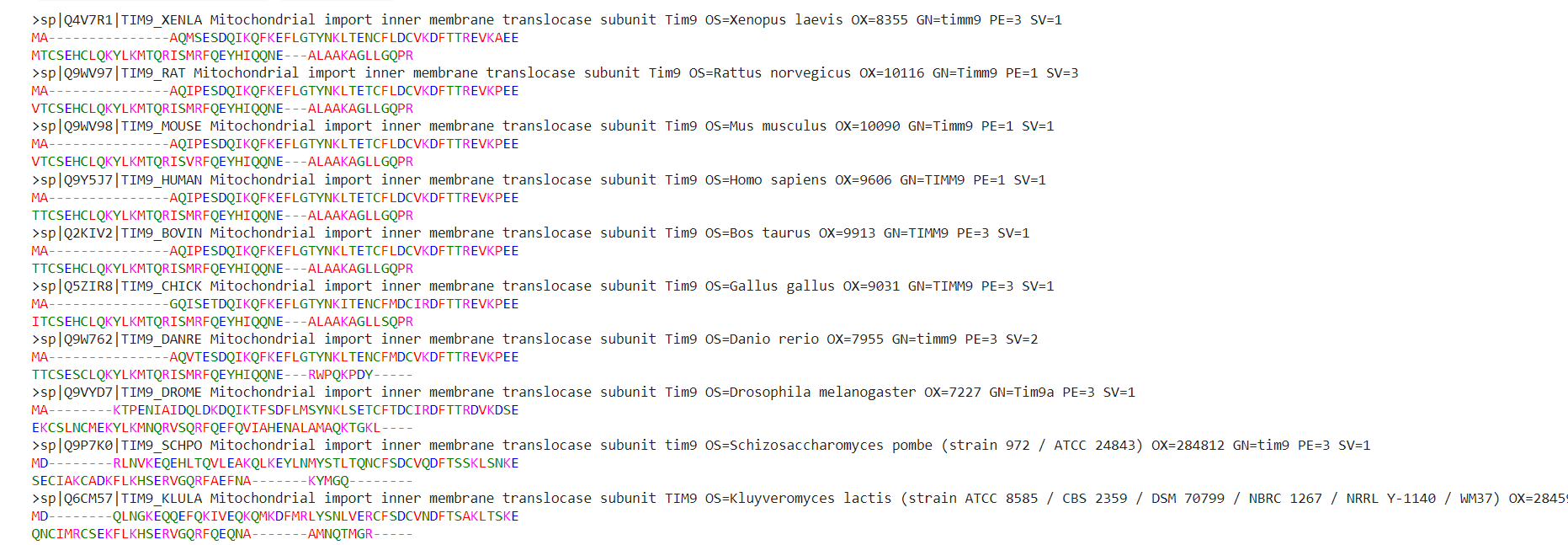


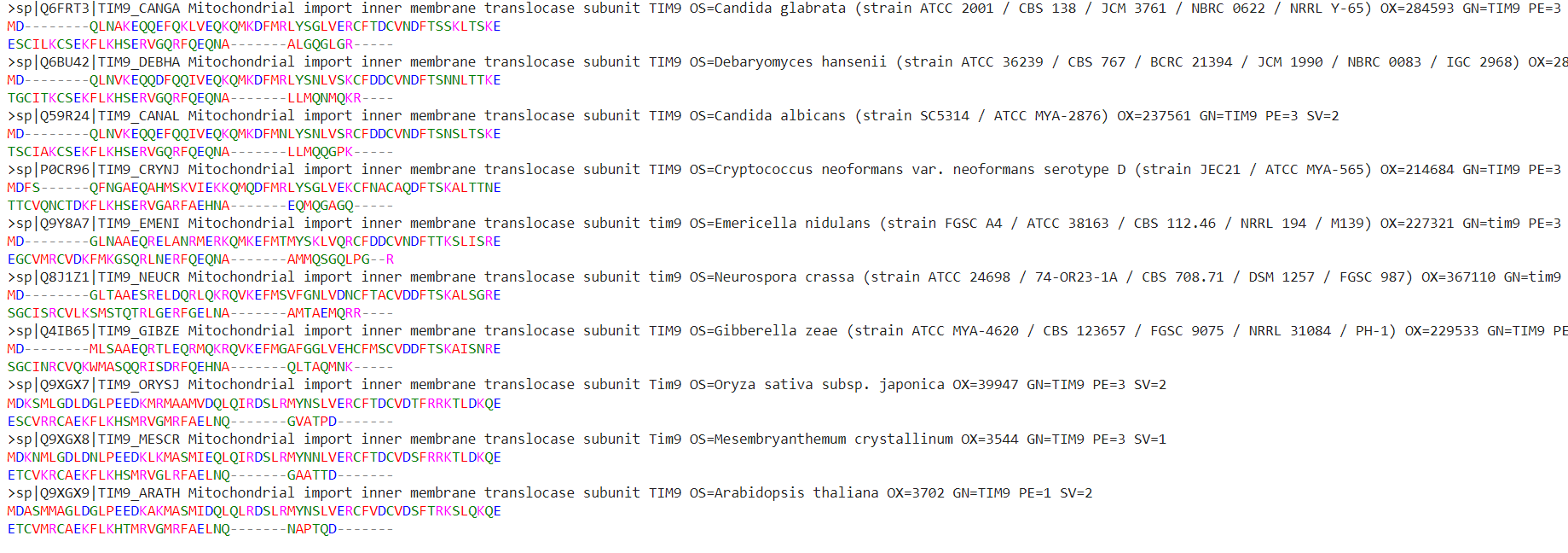
**Clustal Omega:**

****

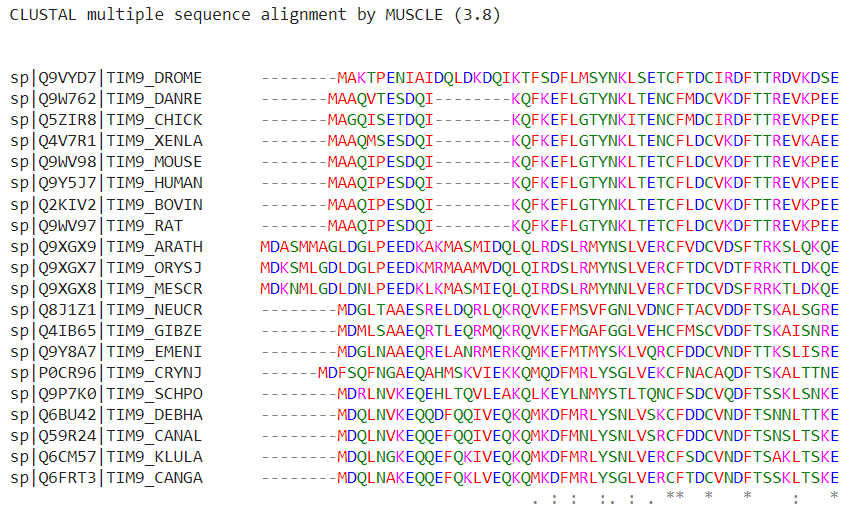
****

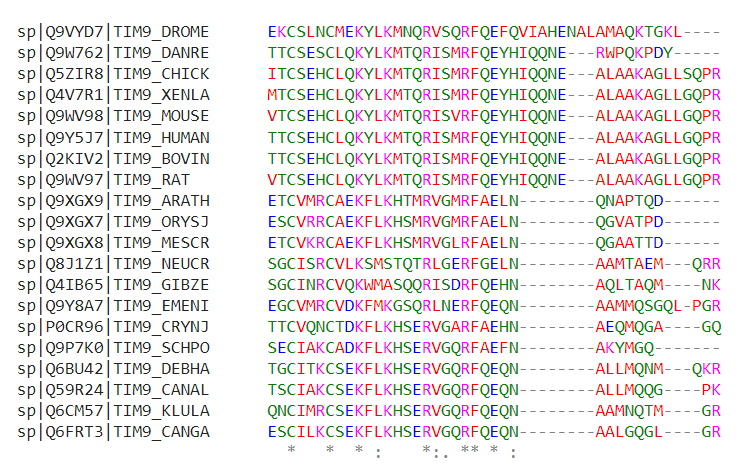
**MAFFT:**

****



**MUSCLE:**

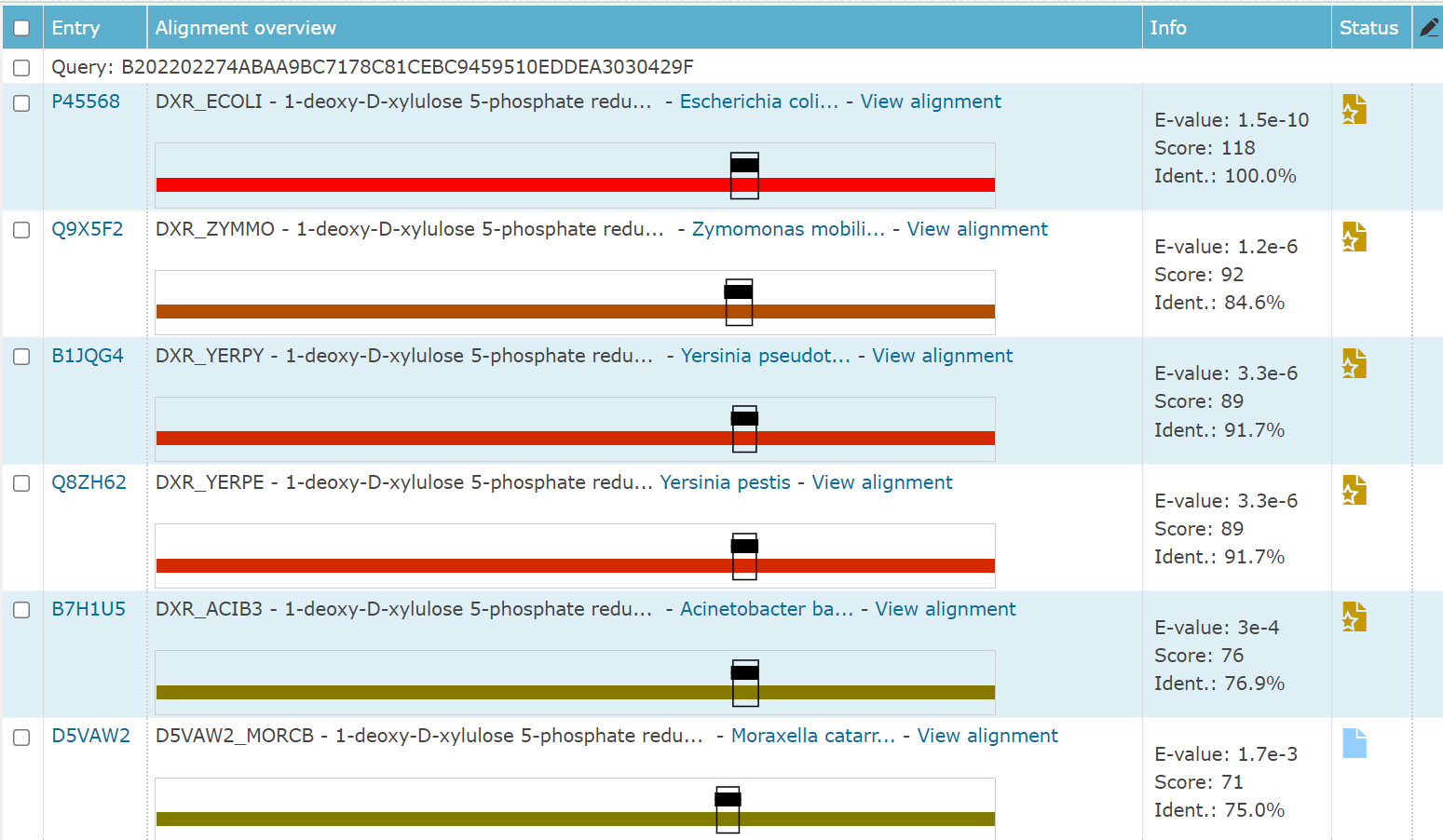
****

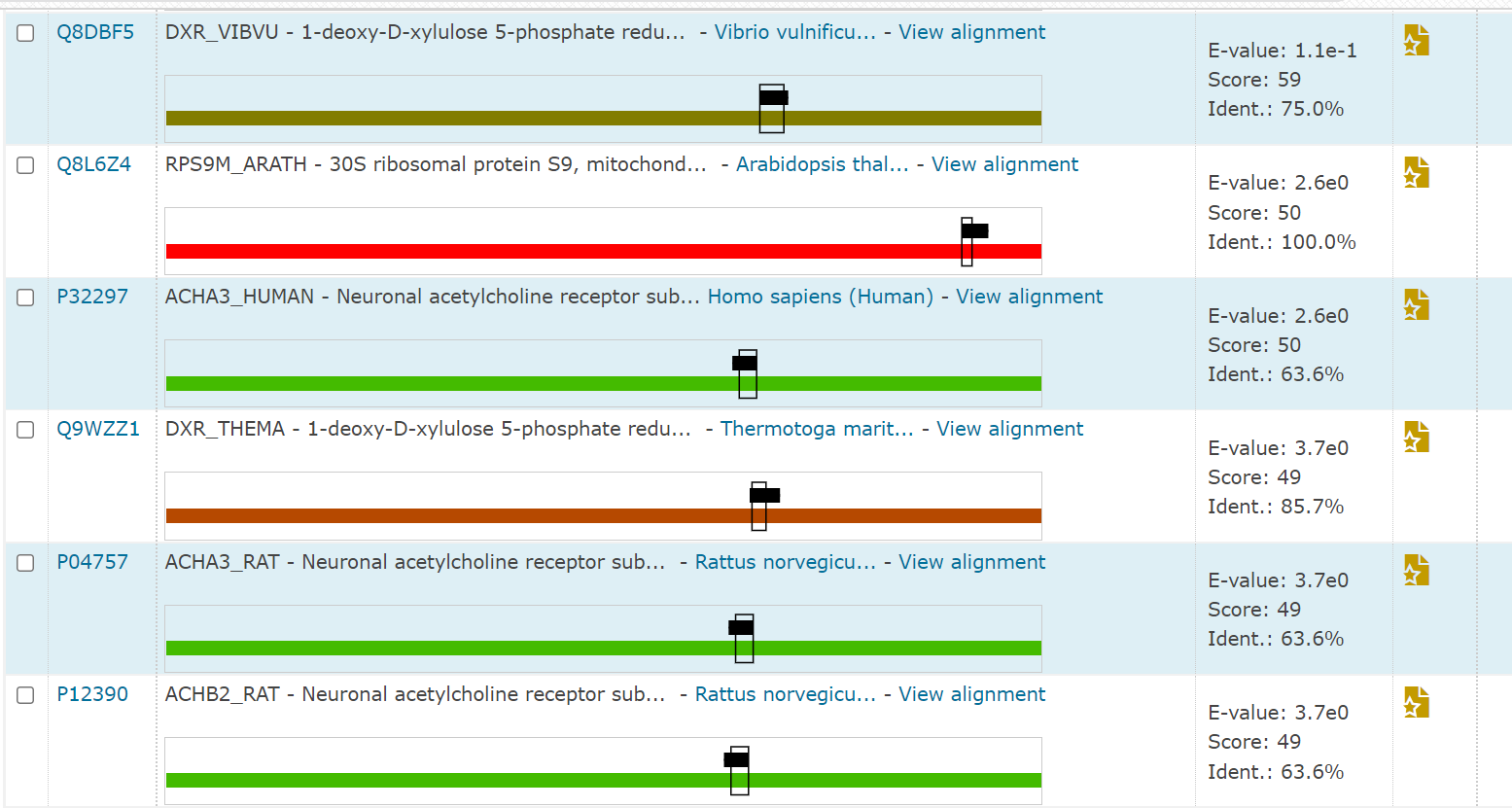
****

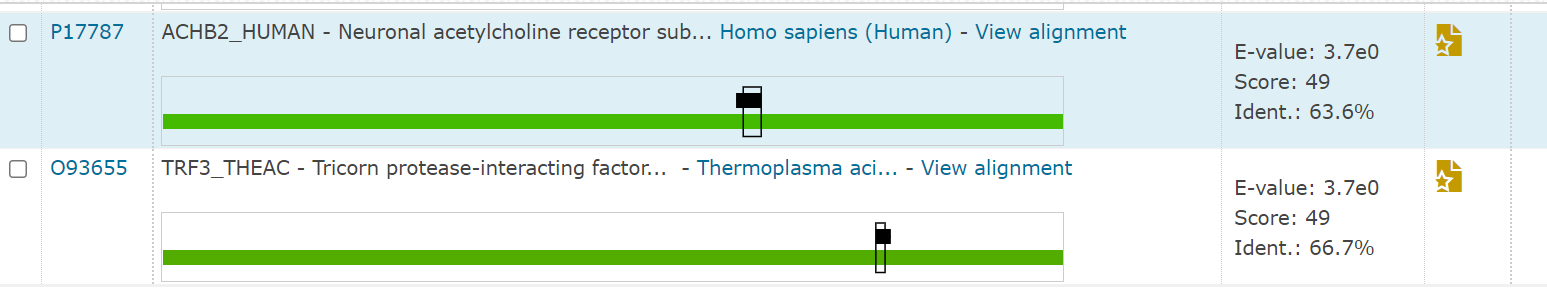
Five residue positions that are aligned differently in these methods are:

1. The first residue (M) in all sequences
2. 11 th position (D)
3. The last residue (R)
4. 12 th position (Q)
5. 13th position (L)

8) When blasted the given sequence 'EPDMRTPIAHTMAW', with the PDB database the results are:







A total of 14 hits were found with the PDB database.

The most significance of the results is that it has 100% sequence identity with two entries **P45568,** which is a reductoisomerase in E.coli strain and **Q8L6Z4,** which is a ribosomal protein in *Arabidopsis thaliana*. Also, it has two more sequences with sequence identity of more than 90.