BIOINFORMATICS
Practical 5

Vineet Patel
BS19B033

Question 1:

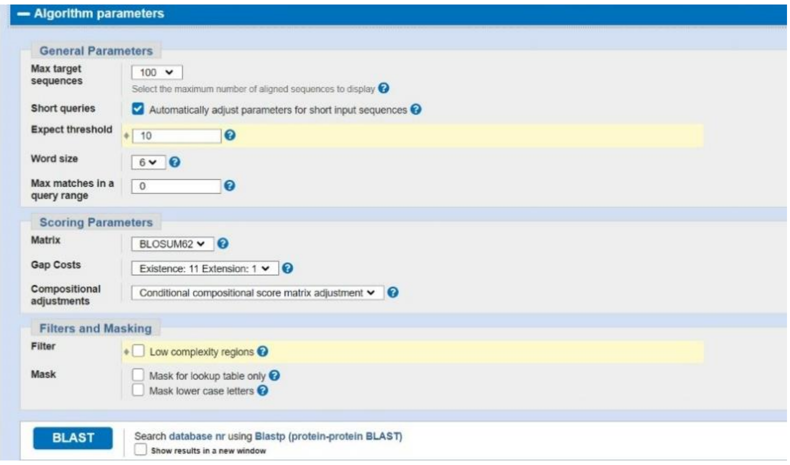
Analysis – “nr” database, of the given 100 results,

- E value is 0.0 for all the results
- There are **23** results that have **100%** Query coverage
- There is **1** result with **100%** identity
- **66.96** is the Lowest percentage identity observed Lowest percentage identity observed

Analysis - “Swiss_PROT” database, of the given 100 results,

- Only 1 E-value is 0.0. After that, E-value increases until it finally reaches **8.4** for one result.
- Query coverage maximum value is **98%** (only 1 results) and goes to min of **6%**.
- No result has 100% identity.
- **22.07%** is the Lowest percentage identity observed Lowest percentage identity observed

Question 2:



— Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 6

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: ☐ Low complexity regions

Mask: ☐ Mask for lookup table only
☐ Mask lower case letters

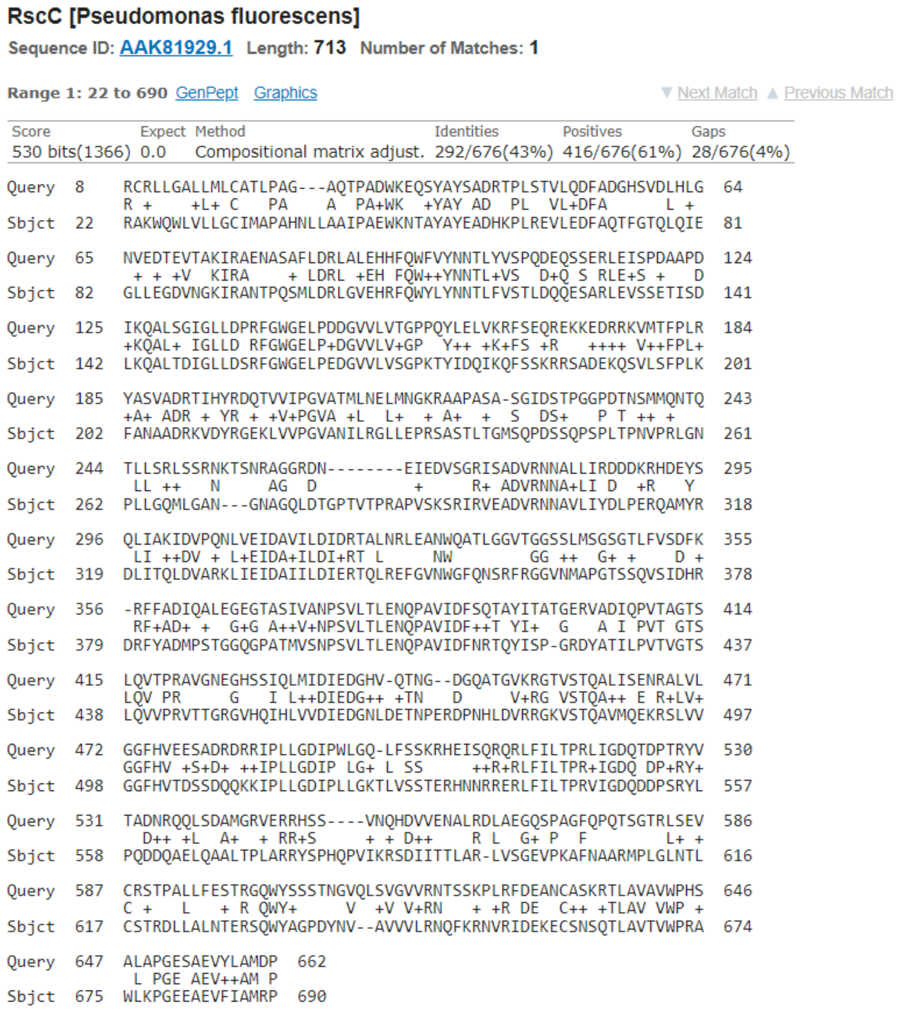
Search database nr using Blastp (protein-protein BLAST)
☐ Show results in a new window

General parameters displayed in comparison:
1. Max target sequences
2. Expected threshold
3. Word size
4. Maximum matches in a query range

Scoring parameters –
1. Matrix
2. Gap costs
3. Compositional alignments

Filter and Masking –
1. Filter options
2. Mask options

Question 3:



RscC [Pseudomonas fluorescens]

Sequence ID: AAK81929.1 Length: 713 Number of Matches: 1

Range 1: 22 to 690 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
530 bits(1366)	0.0	Compositional matrix adjust.	292/676(43%)	416/676(61%)	28/676(4%)

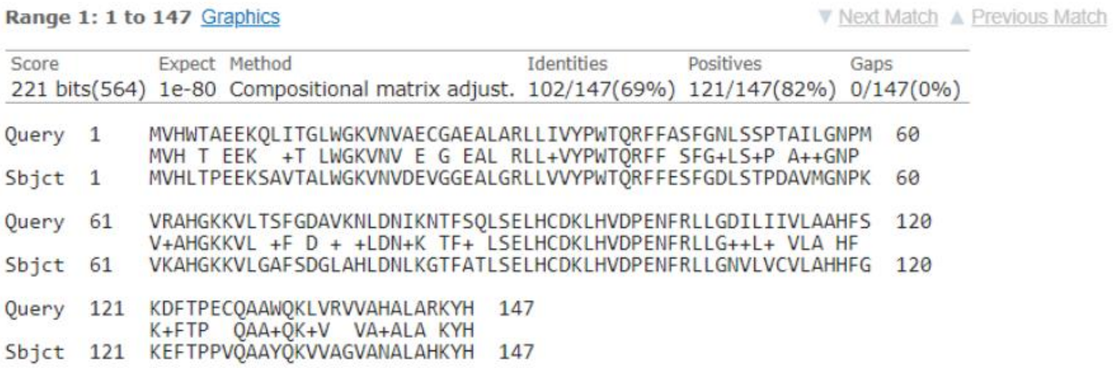
```
Query  8    RCRLLGALLMLCATLPAG---AQTPADWKEQSYAYSADRTPLSTVLQDFADGHSVDLHLG  64
            R +    +L+ C   PA    A  PA+WK  +YAY AD  PL  VL+DFA      L +
Sbjct  22   RAKWQWLVLLGCIMAPAHNLLAAIPAEWKNTAYAYEADHKPLREVLEDFAQTFGTQLQIE  81

Query  65   NVEDTEVTAKIRAENASAFLDRLALEHHFQWFVYNNTLYVSPQDEQSSERLEISPDAAPD  124
             + + +V  KIRA    + LDRL +EH FQW++YNNTL+VS  D+Q S RLE+S +   D
Sbjct  82   GLLEGDVNGKIRANTPQSMLDRLGVEHRFQWYLYNNTLFVSTLDQQESARLEVSSETISD  141

Query  125  IKQALSGIGLLDPRFGWGELPDDGVVLVTGPPQYLELVKRFSEQREKKEDRRKVMTFPLR  184
            +KQAL+ IGLLD RFGWGELP+DGVVLV+GP  Y++ +K+FS +R   ++++ V++FPL+
Sbjct  142  LKQALTDIGLLDSRFGWGELPEDGVVLVSGPKTYIDQIKQFSSKRRSADEKQSVLSFPLK  201

Query  185  YASVADRTIHYRDQTVVIPGVATMLNELMNGKRAAPASA-SGIDSTPGGPDTNSMMQNTQ  243
            +A+ ADR + YR + +V+PGVA +L  L+  + A+  +  S  DS+   P T ++ +
Sbjct  202  FANAADRKVDYRGEKLVVPGVANILRGLLEPRSASTLTGMSQPDSSQPSPLTPNVPRLGN  261

Query  244  TLLSRLSSRNKTSNRAGGRDN--------EIEDVSGRISADVRNNALLIRDDDKRHDEYS  295
             LL ++   N     AG  D          +     R+ ADVRNNA+LI D  +R   Y
Sbjct  262  PLLGQMLGAN---GNAGQLDTGPTVTPRAPVSKSRIRVEADVRNNAVLIYDLPERQAMYR  318

Query  296  QLIAKIDVPQNLVEIDAVILDIDRTALNRLEANWQATLGGVTGGSSLMSGSGTLFVSDFK  355
             LI ++DV + L+EIDA+ILDI+RT L     NW        GG ++  G+ +    D +
Sbjct  319  DLITQLDVARKLIEIDAIILDIERTQLREFGVNWGFQNSRFRGGVNMAPGTSSQVSIDHR  378

Query  356  -RFFADIQALEGEGTASIVANPSVLTLENQPAVIDFSQTAYITATGERVADIQPVTAGTS  414
             RF+AD+ +  G+G A++V+NPSVLTLENQPAVIDF++T YI+  G   A I PVT GTS
Sbjct  379  DRFYADMPSTGGQGPATMVSNPSVLTLENQPAVIDFNRTQYISP-GRDYATILPVTVGTS  437

Query  415  LQVTPRAVGNEGHSSIQLMIDIEDGHV-QTNG--DGQATGVKRGTVSTQALISENRALVL  471
            LQV PR     G   I L++DIEDG++ +TN   D     V+RG VSTQA++ E R+LV+
Sbjct  438  LQVVPRVTTGRGVHQIHLVVDIEDGNLDETNPERDPNHLDVRRGKVSTQAVMQEKRSLVV  497

Query  472  GGFHVEESADRDRRIPLLGDIPWLGQ-LFSSKRHEISQRQRLFILTPRLIGDQTDPTRYV  530
            GGFHV +S+D+ ++IPLLGDIP LG+ L SS     ++R+RLFILTPR+IGDQ DP+RY+
Sbjct  498  GGFHVTDSSDQQKKIPLLGDIPLLGKTLVSSTERHNNRRERLFILTPRVIGDQDDPSRYL  557

Query  531  TADNRQQLSDAMGRVERRHSS----VNQHDVVENALRDLAEGQSPAGFQPQTSGTRLSEV  586
              D++ +L  A+  + RR+S     + + D++    R L  G+ P  F        L+ +
Sbjct  558  PQDDQAELQAALTPLARRYSPHQPVIKRSDIITTLAR-LVSGEVPKAFNAARMPLGLNTL  616

Query  587  CRSTPALLFESTRGQWYSSSTNGVQLSVGVVRNTSSKPLRFDEANCASKRTLAVAVWPHS  646
            C +   L   + R QWY+     V  +V V+RN   + +R DE  C++ +TLAV VWP +
Sbjct  617  CSTRDLLALNTERSQWYAGPDYNV--AVVVLRNQFKRNVRIDEKECSNSQTLAVTVWPRA  674

Query  647  ALAPGESAEVYLAMDP  662
             L PGE AEV++AM P
Sbjct  675  WLKPGEEAEVFIAMRP  690
```

Algorithm:

- In blastp, enter the accession number in the first box, then select "Align 2 or multiple sequences" from the drop-down menu.
- Write the accession number of the second one there. Select BLAST now.

Result – 43.20%

Question 4:



Range 1: 1 to 147 Graphics

Score	Expect	Method	Identities	Positives	Gaps
221 bits(564)	1e-80	Compositional matrix adjust.	102/147(69%)	121/147(82%)	0/147(0%)

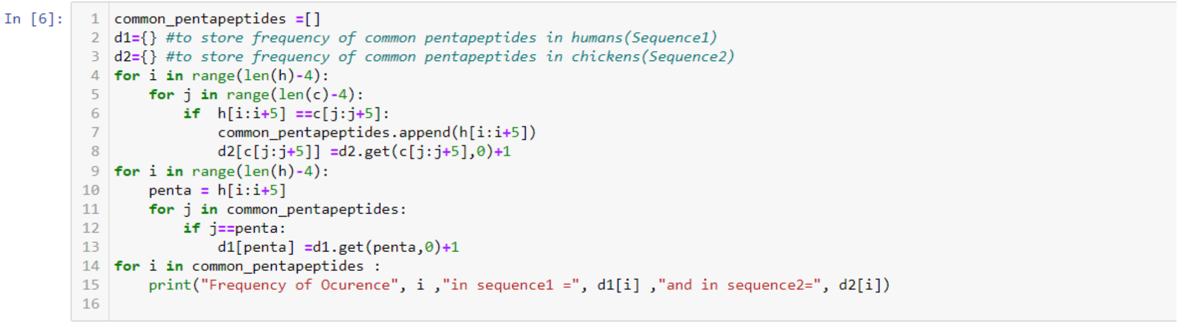
```
Query  1    MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM  60
            MVH T EEK  +T LWGKVNV E G EAL RLL+VYPWTQRFF SFG+LS+P A++GNP
Sbjct  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60

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

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

Algorithm:

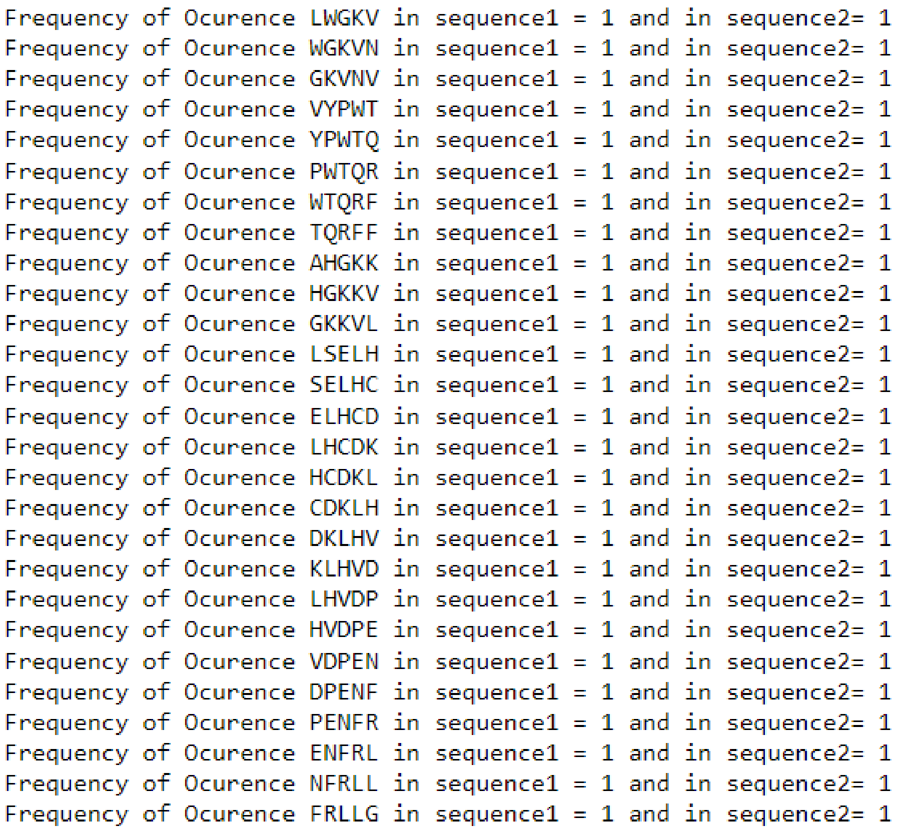
- Obtain both sequences from UniProt.
- BLAST both the sequences and their UniProt IDs, as in the prior query.
- Identity is not the same as similarity. This also considers the nature/properties of two Amino acids.

Question 5:

h(sequence1) = 'MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSD GLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH'
c(sequence2) = 'MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH'



```
In [6]:
common_pentapeptides =[]
d1={} #to store frequency of common pentapeptides in humans(Sequence1)
d2={} #to store frequency of common pentapeptides in chickens(Sequence2)
for i in range(len(h)-4):
    for j in range(len(c)-4):
        if  h[i:i+5] ==c[j:j+5]:
            common_pentapeptides.append(h[i:i+5])
            d2[c[j:j+5]] =d2.get(c[j:j+5],0)+1
for i in range(len(h)-4):
    penta = h[i:i+5]
    for j in common_pentapeptides:
        if j==penta:
            d1[penta] =d1.get(penta,0)+1
for i in common_pentapeptides :
    print("Frequency of Ocurence", i ,"in sequence1 =", d1[i] ,"and in sequence2=", d2[i])
```



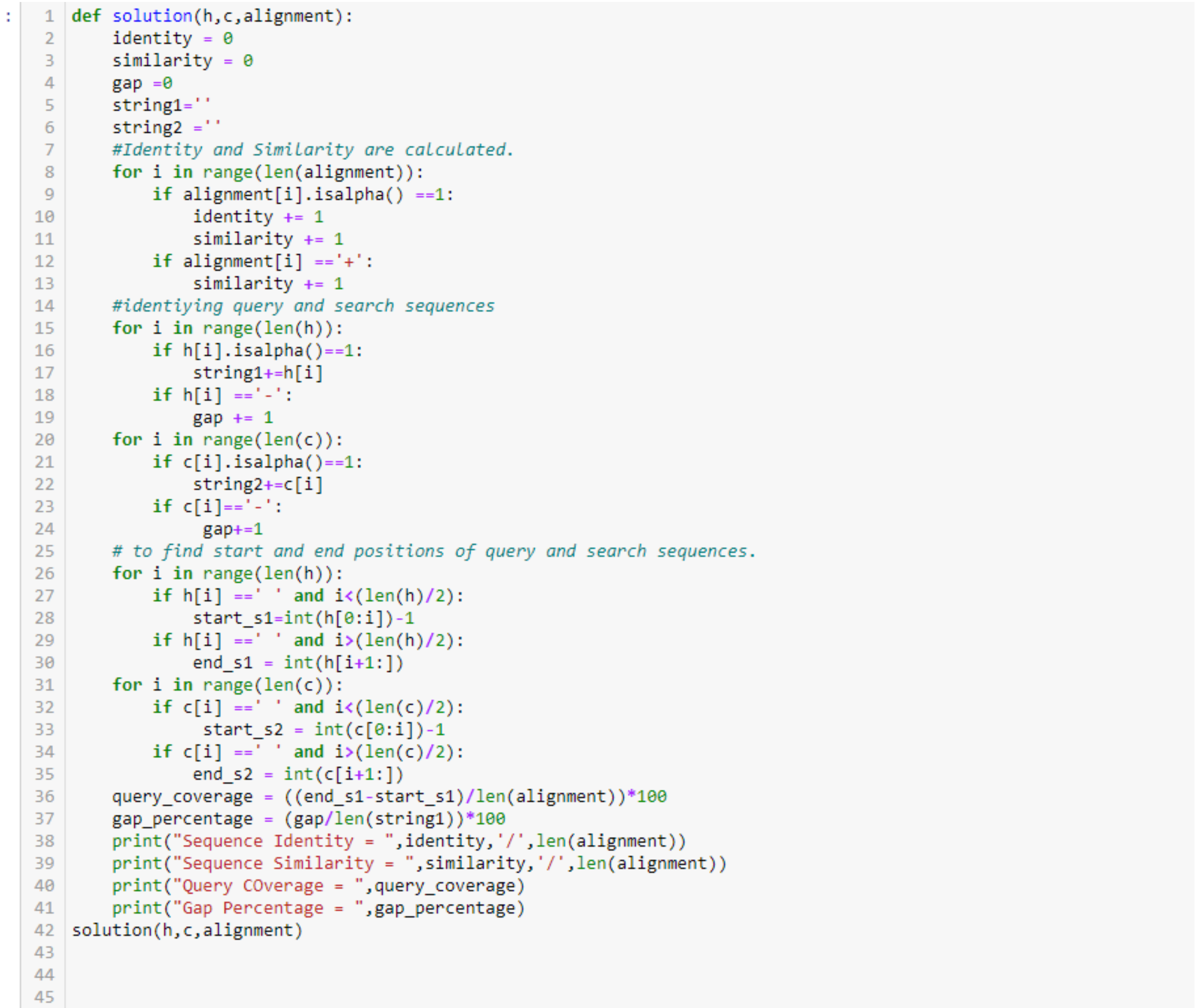
```
Frequency of Ocurence LWGKV in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence WGKVN in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence GKVNV in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence VYPWT in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence YPWTQ in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence PWTQR in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence WTQRF in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence TQRFF in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence AHGKK in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence HGKKV in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence GKKVL in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence LSELH in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence SELHC in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence ELHCD in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence LHCDK in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence HCDKL in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence CDKLH in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence DKLHV in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence KLHVD in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence LHVDP in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence HVDPE in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence VDPEN in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence DPENF in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence PENFR in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence ENFRL in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence NFRLL in sequence1 = 1 and in sequence2= 1
Frequency of Ocurence FRLLG in sequence1 = 1 and in sequence2= 1
```

Question 6:

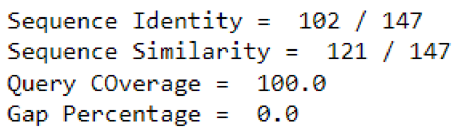
h = '1 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDG LAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147'

c = '1 MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH 147'

alignment = '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'

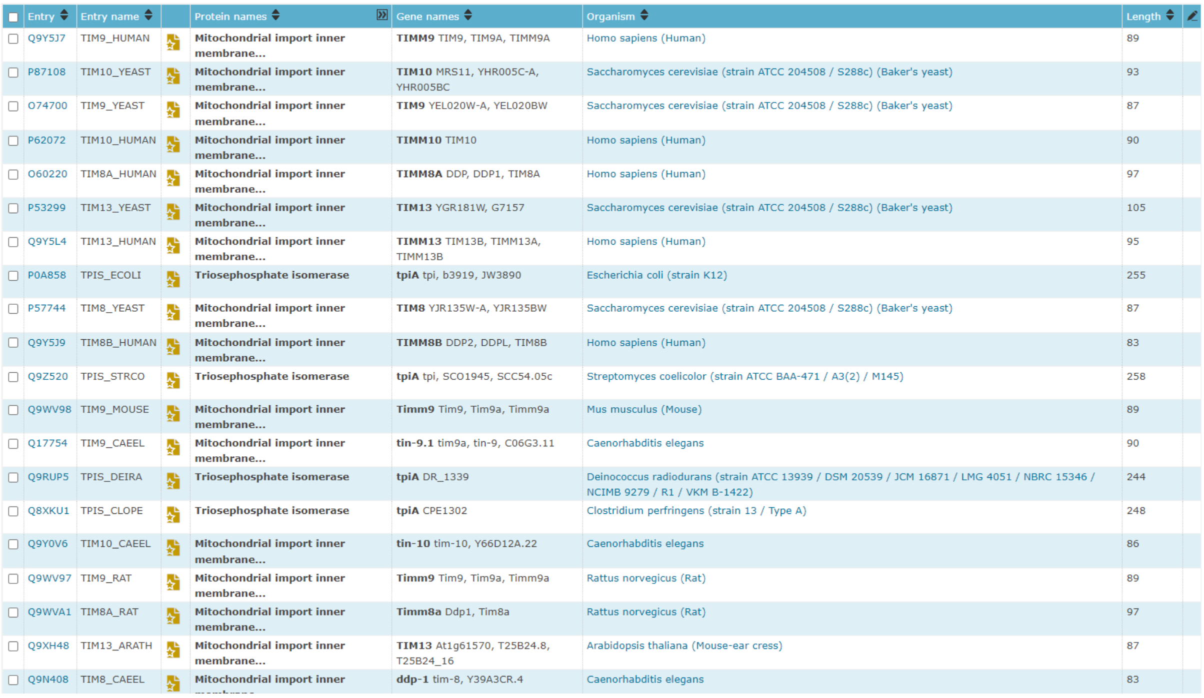


```
def solution(h,c,alignment):
    identity = 0
    similarity = 0
    gap =0
    string1=''
    string2 =''
    #Identity and Similarity are calculated.
    for i in range(len(alignment)):
        if alignment[i].isalpha() ==1:
            identity += 1
            similarity += 1
        if alignment[i] =='+':
            similarity += 1
    #identiying query and search sequences
    for i in range(len(h)):
        if h[i].isalpha()==1:
            string1+=h[i]
        if h[i] =='-':
            gap += 1
    for i in range(len(c)):
        if c[i].isalpha()==1:
            string2+=c[i]
        if c[i]=='-':
             gap+=1
    # to find start and end positions of query and search sequences.
    for i in range(len(h)):
        if h[i] ==' ' and i<(len(h)/2):
            start_s1=int(h[0:i])-1
        if h[i] ==' ' and i>(len(h)/2):
            end_s1 = int(h[i+1:])
    for i in range(len(c)):
        if c[i] ==' ' and i<(len(c)/2):
             start_s2 = int(c[0:i])-1
        if c[i] ==' ' and i>(len(c)/2):
            end_s2 = int(c[i+1:])
    query_coverage = ((end_s1-start_s1)/len(alignment))*100
    gap_percentage = (gap/len(string1))*100
    print("Sequence Identity = ",identity,'/',len(alignment))
    print("Sequence Similarity = ",similarity,'/',len(alignment))
    print("Query COverage = ",query_coverage)
    print("Gap Percentage = ",gap_percentage)
solution(h,c,alignment)
```



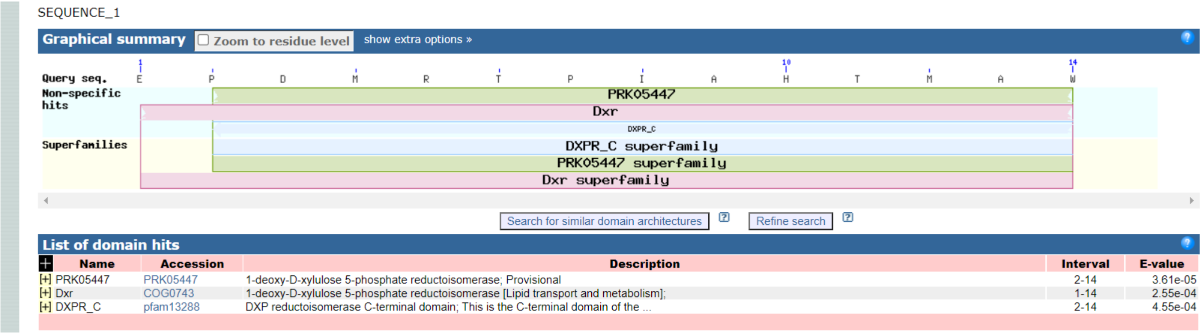
```
Sequence Identity =  102 / 147
Sequence Similarity =  121 / 147
Query COverage =  100.0
Gap Percentage =  0.0
```

Question 7:



□	Entry	Entry name		Protein names	Gene names	Organism	Length
□	Q9Y5J7	TIM9_HUMAN		**Mitochondrial import inner membrane...**	**TIMM9** TIM9, TIM9A, TIMM9A	Homo sapiens (Human)	89
□	P87108	TIM10_YEAST		**Mitochondrial import inner membrane...**	**TIM10** MRS11, YHR005C-A, YHR005BC	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	93
□	O74700	TIM9_YEAST		**Mitochondrial import inner membrane...**	**TIM9** YEL020W-A, YEL020BW	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	87
□	P62072	TIM10_HUMAN		**Mitochondrial import inner membrane...**	**TIMM10** TIM10	Homo sapiens (Human)	90
□	O60220	TIM8A_HUMAN		**Mitochondrial import inner membrane...**	**TIMM8A** DDP, DDP1, TIM8A	Homo sapiens (Human)	97
□	P53299	TIM13_YEAST		**Mitochondrial import inner membrane...**	**TIM13** YGR181W, G7157	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	105
□	Q9Y5L4	TIM13_HUMAN		**Mitochondrial import inner membrane...**	**TIMM13** TIM13B, TIMM13A, TIMM13B	Homo sapiens (Human)	95
□	P0A858	TPIS_ECOLI		**Triosephosphate isomerase**	**tpiA** tpi, b3919, JW3890	Escherichia coli (strain K12)	255
□	P57744	TIM8_YEAST		**Mitochondrial import inner membrane...**	**TIM8** YJR135W-A, YJR135BW	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	87
□	Q9Y5J9	TIM8B_HUMAN		**Mitochondrial import inner membrane...**	**TIMM8B** DDP2, DDPL, TIM8B	Homo sapiens (Human)	83
□	Q9Z520	TPIS_STRCO		**Triosephosphate isomerase**	**tpiA** tpi, SCO1945, SCC54.05c	Streptomyces coelicolor (strain ATCC BAA-471 / A3(2) / M145)	258
□	Q9WV98	TIM9_MOUSE		**Mitochondrial import inner membrane...**	**Timm9** Tim9, Tim9a, Timm9a	Mus musculus (Mouse)	89
□	Q17754	TIM9_CAEEL		**Mitochondrial import inner membrane...**	**tin-9.1** tim9a, tin-9, C06G3.11	Caenorhabditis elegans	90
□	Q9RUP5	TPIS_DEIRA		**Triosephosphate isomerase**	**tpiA** DR_1339	Deinococcus radiodurans (strain ATCC 13939 / DSM 20539 / JCM 16871 / LMG 4051 / NBRC 15346 / NCIMB 9279 / R1 / VKM B-1422)	244
□	Q8XKU1	TPIS_CLOPE		**Triosephosphate isomerase**	**tpiA** CPE1302	Clostridium perfringens (strain 13 / Type A)	248
□	Q9Y0V6	TIM10_CAEEL		**Mitochondrial import inner membrane...**	**tin-10** tim-10, Y66D12A.22	Caenorhabditis elegans	86
□	Q9WV97	TIM9_RAT		**Mitochondrial import inner membrane...**	**Timm9** Tim9, Tim9a, Timm9a	Rattus norvegicus (Rat)	89
□	Q9WVA1	TIM8A_RAT		**Mitochondrial import inner membrane...**	**Timm8a** Ddp1, Tim8a	Rattus norvegicus (Rat)	97
□	Q9XH48	TIM13_ARATH		**Mitochondrial import inner membrane...**	**TIM13** At1g61570, T25B24.8, T25B24_16	Arabidopsis thaliana (Mouse-ear cress)	87
□	Q9N408	TIM8_CAEEL		**Mitochondrial import inner**	**ddp-1** tim-8, Y39A3CR.4	Caenorhabditis elegans	83

Question 8:



SEQUENCE_1

+	Name	Accession	Description	Interval	E-value
[+]	PRK05447	PRK05447	1-deoxy-D-xylulose 5-phosphate reductoisomerase; Provisional	2-14	3.61e-05
[+]	Dxr	COG0743	1-deoxy-D-xylulose 5-phosphate reductoisomerase [Lipid transport and metabolism];	1-14	2.55e-04
[+]	DXPR_C	pfam13288	DXP reductoisomerase C-terminal domain; This is the C-terminal domain of the ...	2-14	4.55e-04

The given sequence is very short. Hence it appears to be a part of a lot of protein families.

Residue 2-14 is a domain named PRK05447 which is conserved across many organisms. BLAST results show that this sequence is very commonly found in – Escherichia Coli, Klebsiella pneumoniae. Of the 100 sequences aligned, about 10% of the sequences have very minimal E-Value. Highest reported E-Value is 28, and this sequence has 57% query coverage and 75% identity with the given sequence