BT3040 – BIOINFORMATICS – Assignment 4

Submitted by Sahana (BE17B038)

Question 1 –

Database – Nucleotide collection (nr/nt) has collated information from all other databases as well, such as UniPROT, Swiss_PROT, TrEMBL, etc. While, Swiss_Prot is just one of the available databases for protein sequences.

Hence, there will be a lot more similar sequences available in "nr" than Swiss_PROT.

The given sequence is a Lysosomal-associated membrane protein from Homo Sapiens.

As mentioned earlier, "nr" is a bigger database. Hence, there will be a lot more results with very low E-value, higher Query Coverage than of the results from "Swiss_PROT".

Analysis –

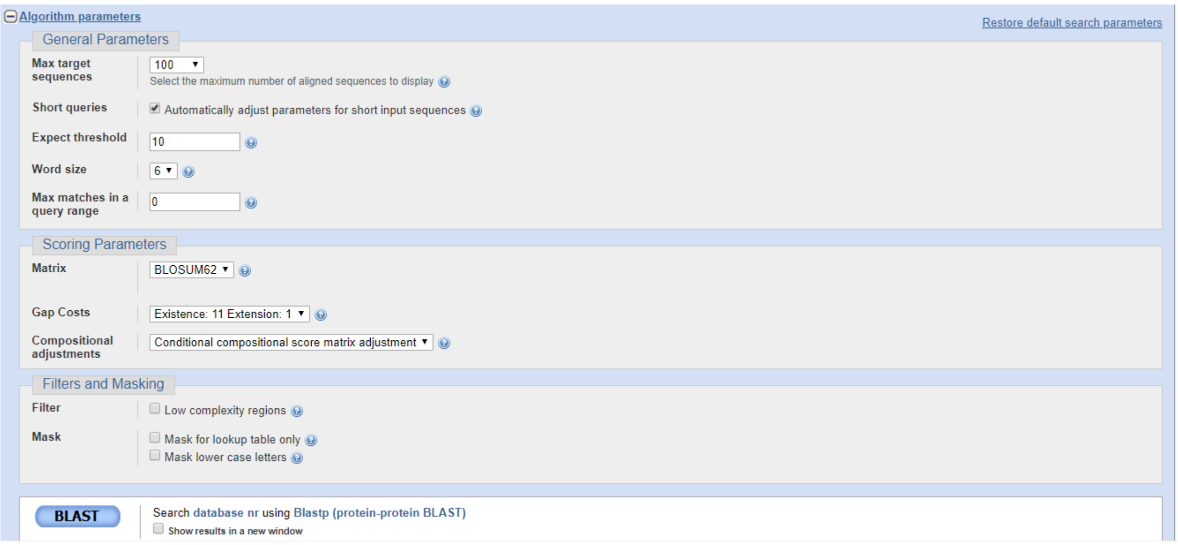
From "nr" database, of the given 100 results,

- all results have close to 0.0 E-value.
- 47 results have 100% Query coverage.
- 3 results with 100% identity
- Lowest percentage identity observed is 71.67% with the lysosome-associated membrane glycoprotein 1 isoform X1 [Urocitellus parryii]

From "Swiss_PROT" database, of the given 100 results,

- Only 2 E-values are 0.0. Thereafter, E-value increases until it finally reaches 5.9 for one result.
- Query coverage also ranges from 100% (only 2 results) to 6%.
- Only 1 result has 100% identity.
- Lowest percentage identity = 23.08% with the RecName: Full=Lysosome-associated membrane glycoprotein 5; AltName: Full=Lysosome-associated membrane protein 5; Short=LAMP-5; Flags: Precursor [Xenopus tropicalis]

Question 2 –



⊟ Algorithm parameters

Restore default search 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

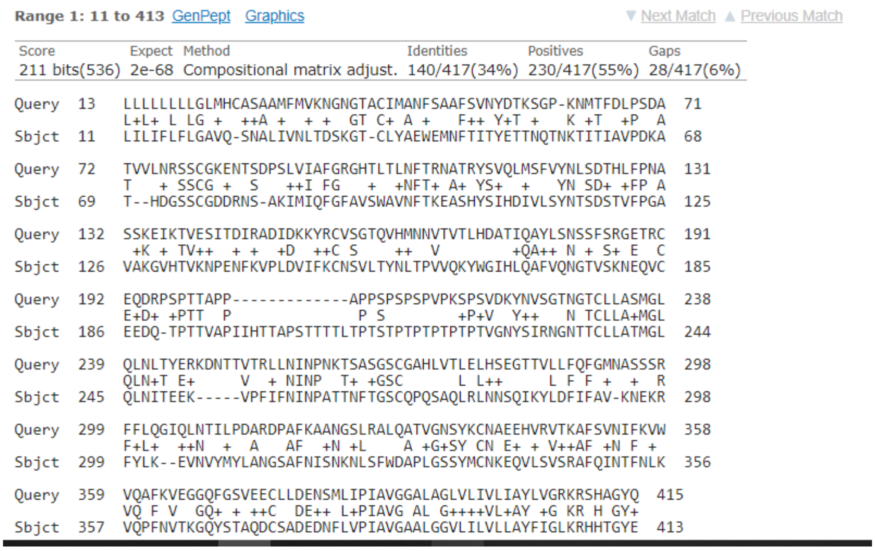
Scoring parameters –

1. Matrix
2. Gap costs
3. Compositional alignments

Filter and Masking –

1. Filter options
2. Mask options

Question 3 –



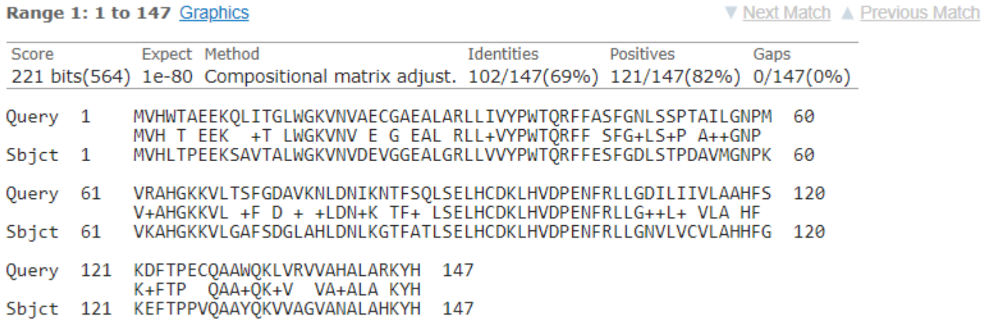
```
Range 1: 11 to 413 GenPept  Graphics                      ▼ Next Match ▲ Previous Match

Score          Expect  Method                          Identities      Positives       Gaps
211 bits(536)  2e-68   Compositional matrix adjust.   140/417(34%)   230/417(55%)   28/417(6%)

Query  13   LLLLLLLLGLMHCASAAMFMVKNGNGTACIMANFSAAFSVNYDTKSGP-KNMTFDLPSDA  71
            L+L+ L LG +  ++A +  + +  GT C+ A +   F++ Y+T +   K +T  +P  A
Sbjct  11   LILIFLFLGAVQ-SNALIVNLTDSKGT-CLYAEWEMNFTITYETTNQTNKTITIAVPDKA  68

Query  72   TVVLNRSSCGKENTSDPSLVIAFGRGHTLTLNFTRNATRYSVQLMSFVYNLSDTHLFPNA  131
            T   + SSCG +  S   ++I FG   +  +NFT+ A+ YS+  +   YN SD+ +FP A
Sbjct  69   T--HDGSSCGDDRNS-AKIMIQFGFAVSWAVNFTKEASHYSIHDIVLSYNTSDSTVFPGA  125

Query  132  SSKEIKTVESITDIRADIDKKYRCVSGTQVHMNNVTVTLHDATIQAYLSNSSFSRGETRC  191
             +K + TV++  + +  +D  ++C S    ++  V        +QA++ N + S+ E  C
Sbjct  126  VAKGVHTVKNPENFKVPLDVIFKCNSVLTYNLTPVVQKYWGIHLQAFVQNGTVSKNEQVC  185

Query  192  EQDRPSPTTAPP-------------APPSPSPSPVPKSPSVDKYNVSGTNGTCLLASMGL  238
            E+D+ +PTT  P              P S        +P+V  Y++    N TCLLA+MGL
Sbjct  186  EEDQ-TPTTVAPIIHTTAPSTTTTLTPTSTPTPTPTPTPTVGNYSIRNGNTTCLLATMGL  244

Query  239  QLNLTYERKDNTTVTRLLNINPNKTSASGSCGAHLVTLELHSEGTTVLLFQFGMNASSSR  298
            QLN+T E+     V  + NINP  T+ +GSC      L L++     L F F +  +  R
Sbjct  245  QLNITEEK-----VPFIFNINPATTNFTGSCQPQSAQLRLNNSQIKYLDFIFAV-KNEKR  298

Query  299  FFLQGIQLNTILPDARDPAFKAANGSLRALQATVGNSYKCNAEEHVRVTKAFSVNIFKVW  358
            F+L+  ++N  +  A   AF  +N +L    A +G+SY CN E+ + V++AF +N F +
Sbjct  299  FYLK--EVNVYMYLANGSAFNISNKNLSFWDAPLGSSYMCNKEQVLSVSRAFQINTFNLK  356

Query  359  VQAFKVEGGQFGSVEECLLDENSMLIPIAVGGALAGLVLIVLIAYLVGRKRSHAGYQ  415
            VQ F V  GQ+ + ++C  DE++ L+PIAVG AL G++++VL+AY +G KR H GY+
Sbjct  357  VQPFNVTKGQYSTAQDCSADEDNFLVPIAVGAALGGVLILVLLAYFIGLKRHHTGYE  413
```

Algorithm -

1. In blastp – put accession number in the first box, and then choose "Align 2 or multiple sequences"
2. Write second one's accession number there. Now, select BLAST.

Result – 33.57%

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 –

1. Go to UniProt and get both the sequences.
2. Similar to the previous question, BLAST both the sequences or their UniProt IDs.
3. Similarity is not identity. This also takes into account the similar nature/properties of two amino acids.

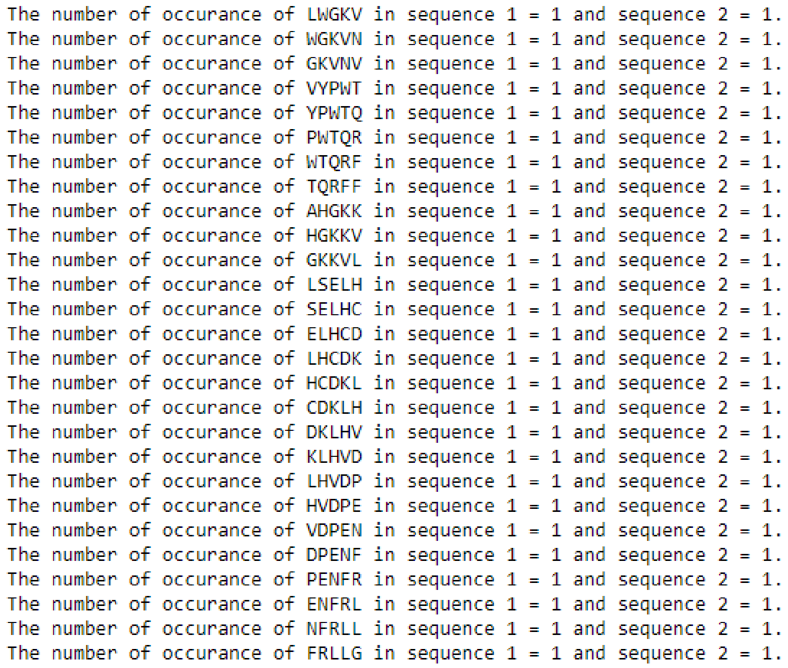
Result – 82% https://blast.ncbi.nlm.nih.gov/Blast.cgi#dtr_Query_4225

Question 5 –

```
Code -
def penta_match(string1, string2):
    n = len(string1)
    m = len(string2)
    occ1 = [0]*(n-4)
    occ2 = [0]*(m-4)
    for i in range(n-4):
        penta = string1[i:i+5]
        for j in range(m-4):
            if string2[j:j+5]==penta:
                occ2[i]+=1
        for k in range(n-4):
            if string1[k:k+5]==penta:
                occ1[i]+=1

    for k in range(len(occ1)):
        if occ1[k]>=1 and occ2[k]>=1:
            print('The number of occurance of %s in sequence 1 = %d and 
sequence 2 = %d.' %(string1[k:k+5],occ1[k],occ2[k]))

Running the code-
```

```
h =
'MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSD
GLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH'
c =
'MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGD
AVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH'
penta_match(h,c)
```

Output-



```
The number of occurance of LWGKV in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of WGKVN in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of GKVNV in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of VYPWT in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of YPWTQ in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of PWTQR in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of WTQRF in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of TQRFF in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of AHGKK in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of HGKKV in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of GKKVL in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of LSELH in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of SELHC in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of ELHCD in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of LHCDK in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of HCDKL in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of CDKLH in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of DKLHV in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of KLHVD in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of LHVDP in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of HVDPE in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of VDPEN in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of DPENF in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of PENFR in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of ENFRL in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of NFRLL in sequence 1 = 1 and sequence 2 = 1.
The number of occurance of FRLLG in sequence 1 = 1 and sequence 2 = 1.
```

Question 6 –

Algorithm –

There are 3 input arguments. The first sequence in alignment – query, second sequence - alignment, third sequence – search. The first and third strings also contain their respective start and stop positions in their actual protein sequences.

Sequence identity = Identical AA residues are denoted by the alphabet of the residue itself in the alignment. Hence the count of all alphabets is taken.

Sequence similarity = This is denoted by a '+' sign, but is actually the sum of occurrences of identical and similar AA residues. This is also summer in the same loop.

Query coverage = The length of query sequence covered while alignment. This is represented by the formula = (aligned_length)/(total_query_length)*100.

Aligned length is given by subtracting the start and end of query sequence obtained from input parameter. Total length of query sequence can also be obtained similarly.

Gap percentage = The total number of gaps in both the query and the search sequence, divided by the total length of aligned sequences.

Code-

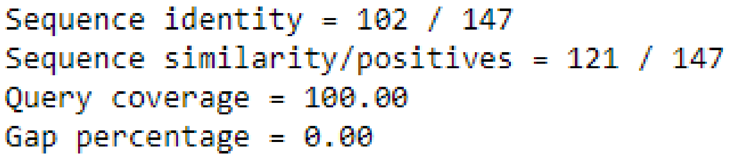
```
def blast_ppty(str1,str2,al):

    al_l = len(al)
    n = len(str1)
    s1 = ''
    m = len(str2)
    s2 = ''
    gap = 0
    %identiying query and search sequences' length and also calculating the
gaps simultaneously.
    for i in range(n):
        if str1[i].isalpha()==1:
            s1+=str1[i]
        if str1[i]=='-':
            gap+=1
    %identifying the start and end positions of query and search sequences.
    for i in range(n):
        if str1[i]==' ' and i<(n/2):
            start_s1=int(str1[0:i])-1
        if str1[i]==' ' and i>(n/2):
            end_s1=int(str1[i+1:])

    for i in range(m):
        if str2[i].isalpha()==1:
            s2+=str2[i]
        if str2[i]=='-':
            gap+=1

    for i in range(m):
        if str2[i]==' ' and i<(m/2):
            start_s2=int(str2[0:i])-1
        if str2[i]==' ' and i>(m/2):
            end_s2=int(str2[i+1:])
    %Values of query_coverage and gap_percentage are calculated
    query_cov = ((end_s1-start_s1)/al_l)*100
    gap_per = (gap/len(s1))*100
    %Identity and Similarity are calculated.
    identity = 0
    similarity = 0
    for i in range(al_l):
        if al[i].isalpha()==1:
            identity+=1
            similarity+=1
        if al[i]=='+':
            similarity+=1
    %The values of identity, similarity, query coverage and gap percentage
are printed/
    print('Sequence identity = %d / %d' %(identity,al_l))
    print('Sequence similarity/positives = %d / %d' %(similarity,al_l))
    print('Query coverage = %3.2f' %query_cov)
    print('Gap percentage = %3.2f' %gap)
```

Running the code-

```
h = '1
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDG
LAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
147'
```

```
c = '1
MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDA
VKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH
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'

blast_ppty(h,c,alignment)
```

Output-



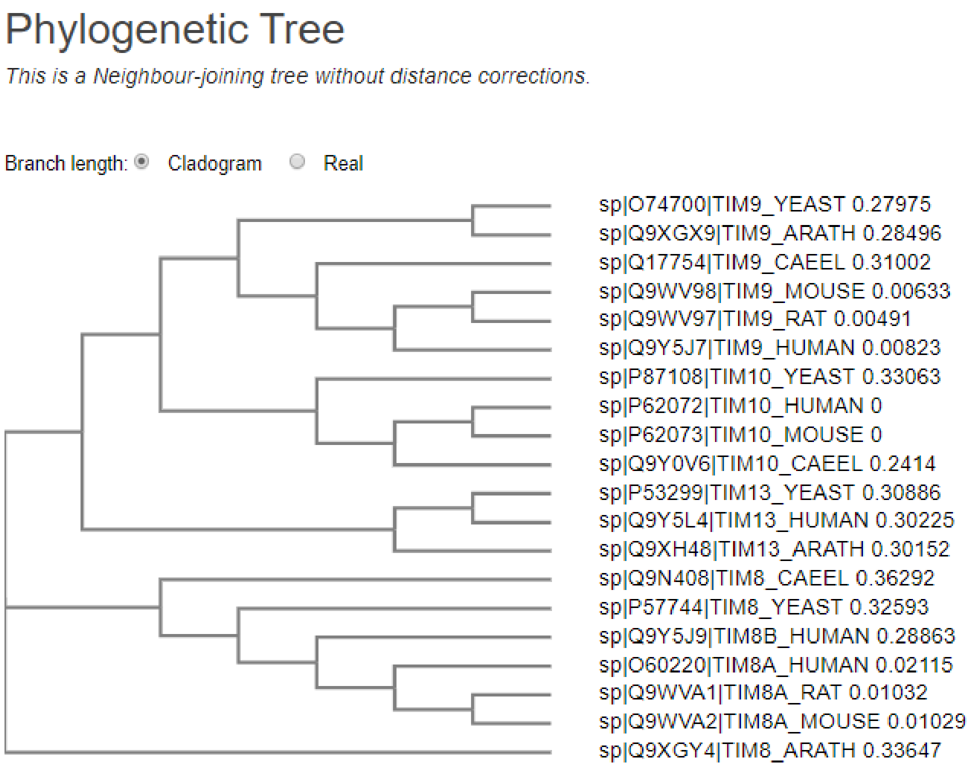
```
Sequence identity = 102 / 147
Sequence similarity/positives = 121 / 147
Query coverage = 100.00
Gap percentage = 0.00
```

Question 7 –

The proteins selected are –

Entry	**Gene names**	**Organism**
Q9Y5J7	TIMM9 TIM9 TIM9A TIMM9A	Homo sapiens (Human)
P87108	TIM10 MRS11 YHR005C-A YHR005BC	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)
P62072	TIMM10 TIM10	Homo sapiens (Human)
O60220	TIMM8A DDP DDP1 TIM8A	Homo sapiens (Human)
O74700	TIM9 YEL020W-A YEL020BW	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)
P53299	TIM13 YGR181W G7157	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)
Q9Y5L4	TIMM13 TIM13B TIMM13A TIMM13B	Homo sapiens (Human)
P57744	TIM8 YJR135W-A YJR135BW	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)
Q9Y5J9	TIMM8B DDP2 DDPL TIM8B	Homo sapiens (Human)
Q17754	tin-9.1 tim9a tin-9 C06G3.11	Caenorhabditis elegans
Q9WV98	Timm9 Tim9 Tim9a Timm9a	Mus musculus (Mouse)
Q9Y0V6	tin-10 tim-10 Y66D12A.22	Caenorhabditis elegans
Q9WV97	Timm9 Tim9 Tim9a Timm9a	Rattus norvegicus (Rat)
Q9WVA1	Timm8a Ddp1 Tim8a	Rattus norvegicus (Rat)
Q9N408	ddp-1 tim-8 Y39A3CR.4	Caenorhabditis elegans
Q9XH48	TIM13 At1g61570 T25B24.8 T25B24_16	Arabidopsis thaliana (Mouse-ear cress)
Q9XGX9	TIM9 EMB2474 At3g46560 F12A12.80	Arabidopsis thaliana (Mouse-ear cress)
Q9WVA2	Timm8a1 Ddp1 Tim8a Timm8a	Mus musculus (Mouse)
P62073	Timm10 Tim10	Mus musculus (Mouse)
Q9XGY4	TIM8 At5g50810 K7B16.3	Arabidopsis thaliana (Mouse-ear cress)

Clustal Omega –



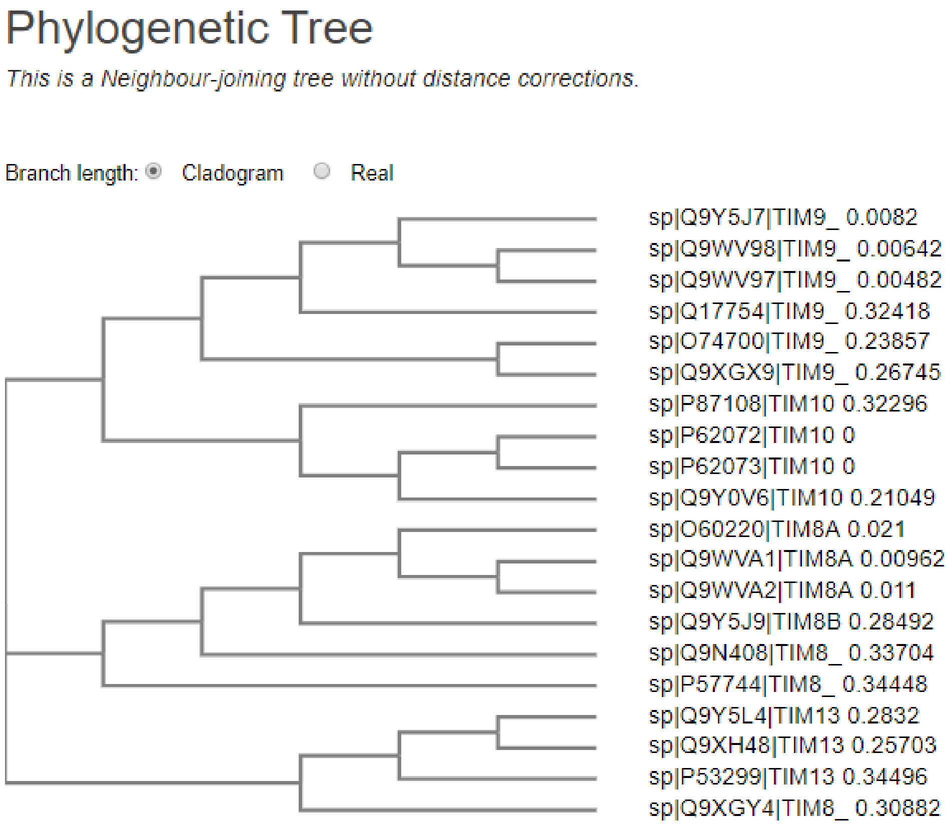
Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

sp|O74700|TIM9_YEAST 0.27975
sp|Q9XGX9|TIM9_ARATH 0.28496
sp|Q17754|TIM9_CAEEL 0.31002
sp|Q9WV98|TIM9_MOUSE 0.00633
sp|Q9WV97|TIM9_RAT 0.00491
sp|Q9Y5J7|TIM9_HUMAN 0.00823
sp|P87108|TIM10_YEAST 0.33063
sp|P62072|TIM10_HUMAN 0
sp|P62073|TIM10_MOUSE 0
sp|Q9Y0V6|TIM10_CAEEL 0.2414
sp|P53299|TIM13_YEAST 0.30886
sp|Q9Y5L4|TIM13_HUMAN 0.30225
sp|Q9XH48|TIM13_ARATH 0.30152
sp|Q9N408|TIM8_CAEEL 0.36292
sp|P57744|TIM8_YEAST 0.32593
sp|Q9Y5J9|TIM8B_HUMAN 0.28863
sp|O60220|TIM8A_HUMAN 0.02115
sp|Q9WVA1|TIM8A_RAT 0.01032
sp|Q9WVA2|TIM8A_MOUSE 0.01029
sp|Q9XGY4|TIM8_ARATH 0.33647

MAFFT –

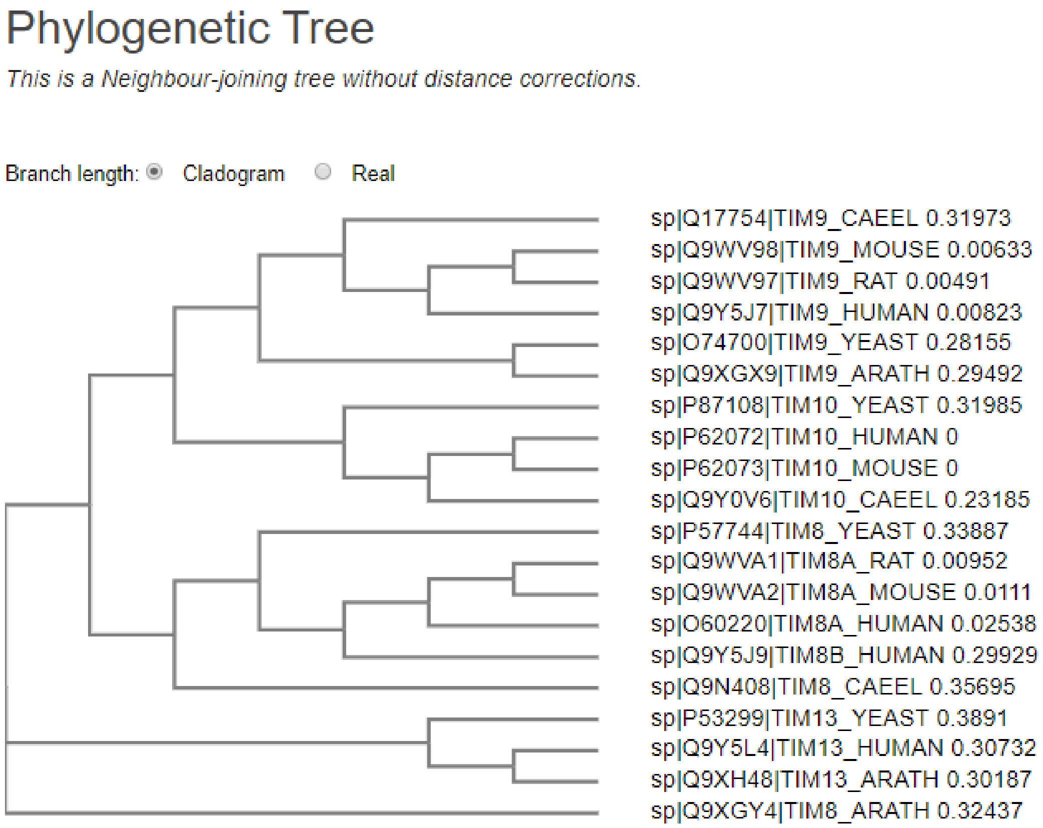


Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

MUSCLE –



Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

sp|Q17754|TIM9_CAEEL 0.31973
sp|Q9WV98|TIM9_MOUSE 0.00633
sp|Q9WV97|TIM9_RAT 0.00491
sp|Q9Y5J7|TIM9_HUMAN 0.00823
sp|O74700|TIM9_YEAST 0.28155
sp|Q9XGX9|TIM9_ARATH 0.29492
sp|P87108|TIM10_YEAST 0.31985
sp|P62072|TIM10_HUMAN 0
sp|P62073|TIM10_MOUSE 0
sp|Q9Y0V6|TIM10_CAEEL 0.23185
sp|P57744|TIM8_YEAST 0.33887
sp|Q9WVA1|TIM8A_RAT 0.00952
sp|Q9WVA2|TIM8A_MOUSE 0.0111
sp|O60220|TIM8A_HUMAN 0.02538
sp|Q9Y5J9|TIM8B_HUMAN 0.29929
sp|Q9N408|TIM8_CAEEL 0.35695
sp|P53299|TIM13_YEAST 0.3891
sp|Q9Y5L4|TIM13_HUMAN 0.30732
sp|Q9XH48|TIM13_ARATH 0.30187
sp|Q9XGY4|TIM8_ARATH 0.32437

Positions of misalignment across three MSA algorithms –

Position 1

- 7 out of 20 alignments start with Methionine in Clustal Omega. Although the remaining 13 sequences also start with 'M', it is not aligned with the rest.
- 17 out of 20 alignments start with 'M' in MAFFT. The remaining 3 do not start at position 1.
- Only 1 sequence starts its alignment at position 1 in MUSCLE. Remaining sequences start elsewhere.

Position 60

- In MAFFT, Cysteine has been aligned through the sequence.
- In Clustal Omega, Phenylalanine has been majorly aligned among the sequences. The cysteine match comes at position 59.
- In MUSCLE, Lysine has been majorly aligned at position 60. Cysteine alignment comes in position 57.

Position 16

- In MAFFT, AA residue is mostly aliphatic polar such as A, V, L, M.

- In Clustal Omega, position 16 is majorly not aligned, except for Glutamine in 1 sequence.
- In MUSCLE, AA residue at position 16 is Serine with probability close to 0.5. In other cases, they are not aligned.

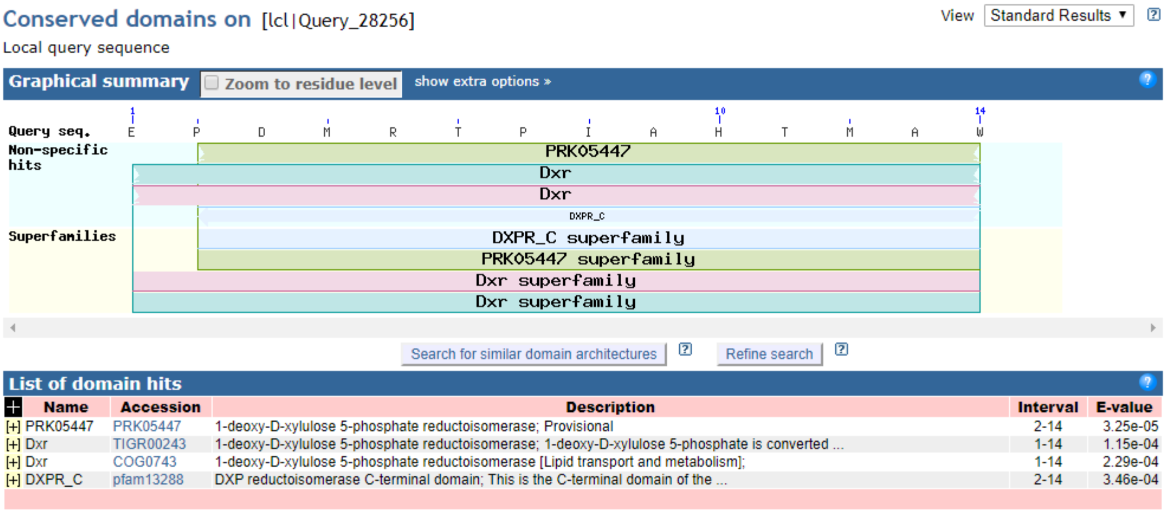
Position 75

- In Clustal Omega, this position is mostly N/G/S/K.
- In MAFFT, this position is mostly occupied by aliphatic polar residue such as V/L.
- In MUSCLE, this position is majorly occupied by E – Glutamic Acid.

Position 61 –

- In clustal omega, this position is mostly occupied by E/D/K/L.
- In MAFFT, Phenylalanine and Tryptophan majorly occupy this position.
- In MUSCLE, this position is Cysteine across all the sequences.

Question 8 –



Conserved domains on [lcl|Query_28256]

Local query sequence

List of domain hits

⊞ Name	Accession	Description	Interval	E-value
[+] PRK05447	PRK05447	1-deoxy-D-xylulose 5-phosphate reductoisomerase; Provisional	2-14	3.25e-05
[+] Dxr	TIGR00243	1-deoxy-D-xylulose 5-phosphate reductoisomerase; 1-deoxy-D-xylulose 5-phosphate is converted ...	1-14	1.15e-04
[+] Dxr	COG0743	1-deoxy-D-xylulose 5-phosphate reductoisomerase [Lipid transport and metabolism];	1-14	2.29e-04
[+] DXPR_C	pfam13288	DXP reductoisomerase C-terminal domain; This is the C-terminal domain of the ...	2-14	3.46e-04

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

1. *Escherichia Coli*
2. *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.