Assignment 5
BS22B009

All the python codes are present here :

https://colab.research.google.com/drive/1Z10-hDsg5gOT9KMfOD6BQV2JtBE1v72Z?usp=sharing

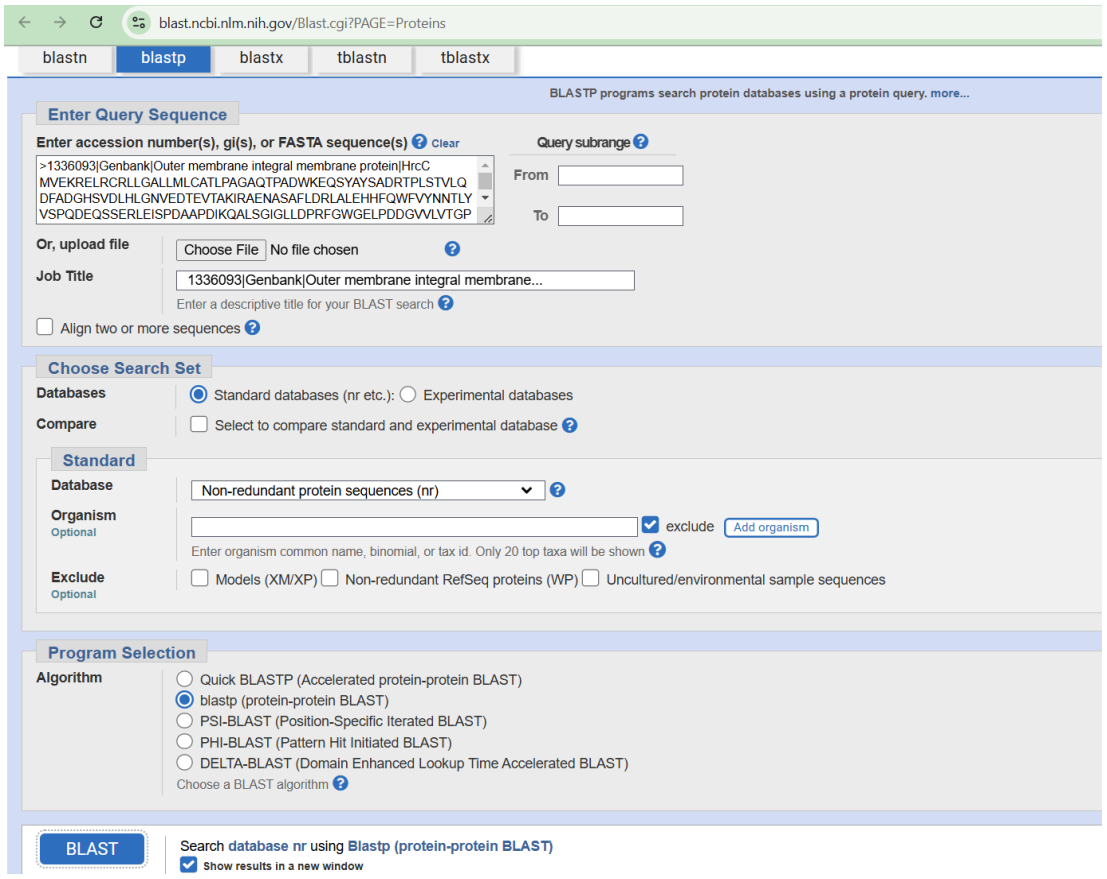
1. Analyze the occurrence of similar proteins in "nr" and SWISS-PROT database for the sequence given below: >1336093|Genbank|Outer membrane integral membrane protein|HrcC

MVEKRELRCRLLGALLMLCATLPAGAQTPADWKEQSYAYSADRTPLSTVLQDFADGHSVD
LHLGNVEDTEVTAKIRAENASAFLDRLALEHHFQWFVYNNTLYVSPQDEQSSERLEISPD
AAPDIKQALSGIGLLDPRFGWGELPDDGVVLVTGPPQYLELVKRFSEQREKKEDRRKVMT
FPLRYASVADRTIHYRDQTVVIPGVATMLNELMNGKRAAPASASGIDSTPGGPDTNSMMQ
NTQTLLSRLSSRNKTSNRAGGRDNEIEDVSGRISADVRNNALLIRDDDKRHDEYSQLIAK
IDVPQNLVEIDAVILDIDRTALNRLEANWQATLGGVTGGSSLMSGSGTLFVSDFKRFFAD
IQALEGEGTASIVANPSVLTLENQPAVIDFSQTAYITATGERVADIQPVTAGTSLQVTPR
AVGNEGHSSIQLMIDIEDGHVQTNGDGQATGVKRGTVSTQALISENRALVLGGFHVEESA
DRDRRIPLLGDIPWLGQLFSSKRHEISQRQRLFILTPRLIGDQTDPTRYVTADNRQQLSD
AMGRVERRHSSVNQHDVVENALRDLAEGQSPAGFQPQTSGTRLSEVCRSTPALLFESTRG
QWYSSSTNGVQLSVGVVRNTSSKPLRFDEANCASKRTLAVAVWPHSALAPGESAEVYLA
M DPSRVLHASRESLLNR

<u>Steps:</u>

Use BLAST (Basic Local Alignment Search Tool):
• Go to NCBI BLASTp (for protein sequences):
https://blast.ncbi.nlm.nih.gov/Blast.cgi
• Select "Protein BLAST (BLASTp)".
• Enter the given sequence in the query box.
• Choose the nr (non-redundant) protein database for a broad search.
• Run the BLAST search and analyze the results (sequence similarity, identity, and alignment).

After running BLASTP with the database as "nr", the following protein sequences are found to be similar to the given sequence:



blast.ncbi.nlm.nih.gov/Blast.cgi

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all 100 sequences selected GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]	Erwinia amylovora	1387	1387	100%	0.0	100.00%	676	WP_004155366.1
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]	Erwinia amylovora	1385	1385	100%	0.0	99.85%	676	WP_168385176.1
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]	Erwinia amylovora	1385	1385	100%	0.0	99.85%	676	WP_168421624.1
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]	Erwinia amylovora	1383	1383	100%	0.0	99.85%	676	WP_400167015.1
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]	Erwinia amylovora	1371	1371	100%	0.0	98.97%	677	WP_004168436.1
Type III secretion system outer membrane pore HrcC [Erwinia amylovora ATCC BAA-2158]	Erwinia amylovo...	1369	1369	100%	0.0	98.82%	677	CBX79367.1
type III secretion system outer membrane ring subunit SctC [Erwinia sp. Ejp617]	Erwinia sp. Ejp617	1347	1347	100%	0.0	96.89%	676	WP_014543268.1
type III secretion system outer membrane ring subunit SctC [Erwinia pyrifoliae]	Erwinia pyrifoliae	1343	1343	100%	0.0	96.75%	676	WP_259816781.1
type III secretion system outer membrane ring subunit SctC [Erwinia pyrifoliae]	Erwinia pyrifoliae	1339	1339	100%	0.0	96.45%	676	WP_012669302.1
HrcC [Erwinia pyrifoliae]	Erwinia pyrifoliae	1337	1337	100%	0.0	96.30%	676	ABA39798.2
type III secretion system outer membrane ring subunit SctC [Erwinia piriflorinigrans]	Erwinia piriflorini...	1269	1269	100%	0.0	92.46%	676	WP_023653761.1
type III secretion system outer membrane ring subunit SctC [Erwinia tasmaniensis]	Erwinia tasmani...	1242	1242	97%	0.0	93.62%	676	WP_012440288.1
type III secretion system outer membrane ring subunit SctC [Erwinia psidii]	Erwinia psidii	1211	1211	100%	0.0	86.67%	677	WP_124231871.1
type III secretion system outer membrane ring subunit SctC [Erwinia tracheiphila]	Erwinia tracheip...	1184	1184	100%	0.0	83.85%	674	WP_233479868.1
MULTISPECIES: type III secretion system outer membrane ring subunit SctC [Pantoea]	Pantoea	1184	1184	100%	0.0	84.09%	679	WP_275222444.1
type III secretion system outer membrane ring subunit SctC [Erwinia tracheiphila]	Erwinia tracheip...	1183	1183	100%	0.0	83.70%	674	WP_016191026.1
type III secretion system outer membrane ring subunit SctC [Pantoea sp. AS142]	Pantoea sp. AS...	1181	1181	100%	0.0	84.09%	679	WP_337013830.1
type III secretion system outer membrane ring subunit SctC [Pantoea vagans]	Pantoea vagans	1180	1180	100%	0.0	83.95%	679	WP_061060948.1
type III secretion system outer membrane ring subunit SctC [Pantoea vagans]	Pantoea vagans	1177	1177	100%	0.0	83.80%	679	WP_336749516.1
type III secretion system outer membrane ring subunit SctC [Pantoea agglomerans]	Pantoea agglom...	1176	1176	100%	0.0	83.80%	679	WP_323169453.1
type III secretion system outer membrane ring subunit SctC [Pantoea agglomerans]	Pantoea agglom...	1176	1176	100%	0.0	83.65%	679	WP_031590814.1

Feedback

Analysis of Nr non reductant protein sequence BLAST Results:

Top Matches with Erwinia amylovora

- The best hits (top 5 sequences) are from Erwinia amylovora, a well-known plant pathogen responsible for fire blight in apples and pears.
- These sequences have 100% Query Coverage and 100% or 99.85% identity, confirming near-exact matches.
- Max Score: 1387-1385
- E-value: 0.0 (Indicating an extremely strong match)

Presence of Other Erwinia Species

- The T3SS outer membrane ring subunit SctC is highly conserved across multiple Erwinia species, including:
    - E. pyrifoliae, E. piriflorinigrans, E. tasmaniensis, E. psidii, E. tracheiphila, E. papayae, and E. mallotivora.
- These hits have 96-86% sequence identity, showing some level of divergence.

Matches with Pantoea Species

- Pantoea species, closely related to Erwinia, also exhibit significant similarity, with scores ranging from 1184 to 1149 and identity between 84-81%.
- Pantoea includes both opportunistic plant pathogens and beneficial endophytes, so this result suggests a common evolutionary origin for the T3SS structure.

All 100 similar sequences have an E value of 0 and 29 sequences have 100% query coverage

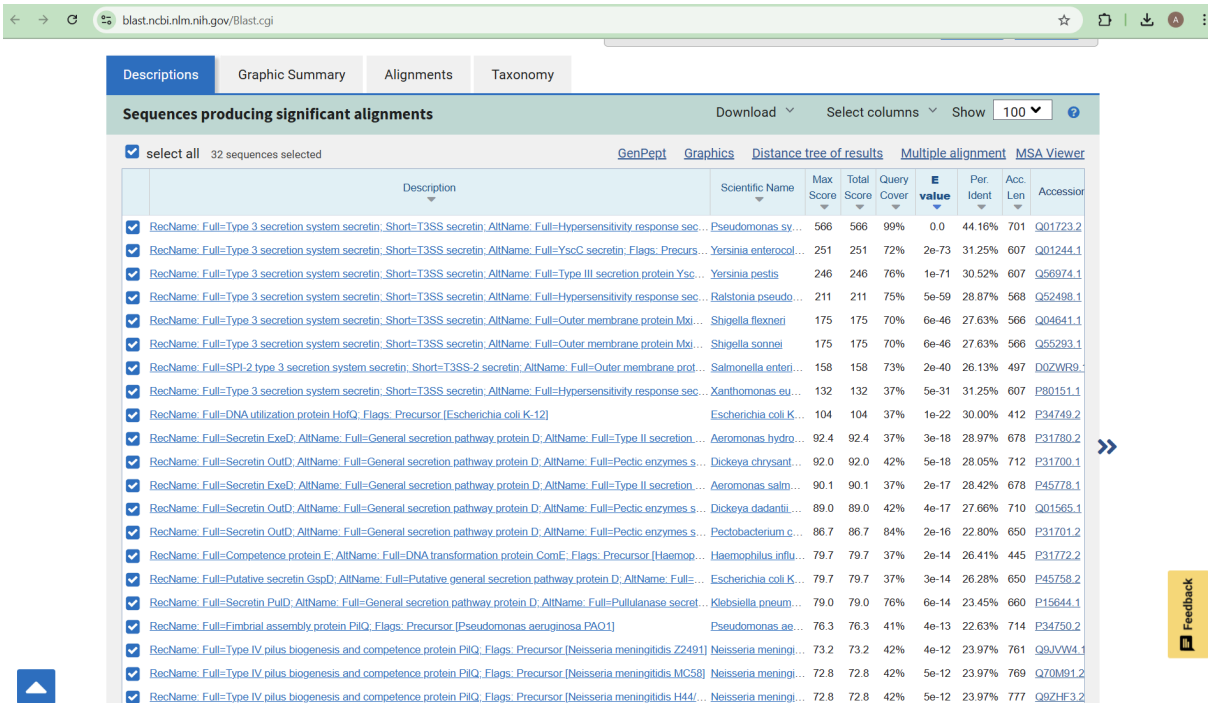
The lowest query coverage is 90% while the lowest percentage identity is 66.96% (type III secretion system outer membrane ring subunit SctC [Dickeya lacustris])

The non-redundant (NR) protein sequence closely matches the T3SS outer membrane ring subunit SctC from *Erwinia amylovora* with 100% identity.

The sequence is highly conserved across multiple Erwinia species and some Pantoea species, confirming its essential function.

Given its role in bacterial pathogenesis, this protein could be a potential target for disease control in plant pathology.

<u>After running BLASTP with the database as “swissprot”, the following protein sequences are found to be similar to the given sequence:</u>



blast.ncbi.nlm.nih.gov/Blast.cgi

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all 32 sequences selected GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Hypersensitivity response sec...	Pseudomonas sy...	566	566	99%	0.0	44.16%	701	Q01723.2
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=YscC secretin; Flags: Precurs...	Yersinia enterocol...	251	251	72%	2e-73	31.25%	607	Q01244.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Type III secretion protein Ysc...	Yersinia pestis	246	246	76%	1e-71	30.52%	607	Q56974.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Hypersensitivity response sec...	Ralstonia pseudo...	211	211	75%	5e-59	28.87%	568	Q52498.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Outer membrane protein Mxi...	Shigella flexneri	175	175	70%	6e-46	27.63%	566	Q04641.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Outer membrane protein Mxi...	Shigella sonnei	175	175	70%	6e-46	27.63%	566	Q55293.1
RecName: Full=SPI-2 type 3 secretion system secretin; Short=T3SS-2 secretin; AltName: Full=Outer membrane prot...	Salmonella enteri...	158	158	73%	2e-40	26.13%	497	D0ZWR9.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Hypersensitivity response sec...	Xanthomonas eu...	132	132	37%	5e-31	31.25%	607	P80151.1
RecName: Full=DNA utilization protein HofQ; Flags: Precursor [Escherichia coli K-12]	Escherichia coli K...	104	104	37%	1e-22	30.00%	412	P34749.2
RecName: Full=Secretin ExeD; AltName: Full=General secretion pathway protein D; AltName: Full=Type II secretion ...	Aeromonas hydro...	92.4	92.4	37%	3e-18	28.97%	678	P31780.2
RecName: Full=Secretin OutD; AltName: Full=General secretion pathway protein D; AltName: Full=Pectic enzymes s...	Dickeya chrysant...	92.0	92.0	42%	5e-18	28.05%	712	P31700.1
RecName: Full=Secretin ExeD; AltName: Full=General secretion pathway protein D; AltName: Full=Type II secretion ...	Aeromonas salm...	90.1	90.1	37%	2e-17	28.42%	678	P45778.1
RecName: Full=Secretin OutD; AltName: Full=General secretion pathway protein D; AltName: Full=Pectic enzymes s...	Dickeya dadantii ...	89.0	89.0	42%	4e-17	27.66%	710	Q01565.1
RecName: Full=Secretin OutD; AltName: Full=General secretion pathway protein D; AltName: Full=Pectic enzymes s...	Pectobacterium c...	86.7	86.7	84%	2e-16	22.80%	650	P31701.2
RecName: Full=Competence protein E; AltName: Full=DNA transformation protein ComE; Flags: Precursor [Haemop...	Haemophilus influ...	79.7	79.7	37%	2e-14	26.41%	445	P31772.2
RecName: Full=Putative secretin GspD; AltName: Full=Putative general secretion pathway protein D; AltName: Full=...	Escherichia coli K...	79.7	79.7	37%	3e-14	26.28%	650	P45758.2
RecName: Full=Secretin PulD; AltName: Full=General secretion pathway protein D; AltName: Full=Pullulanase secret...	Klebsiella pneum...	79.0	79.0	76%	6e-14	23.45%	660	P15644.1
RecName: Full=Fimbrial assembly protein PilQ; Flags: Precursor [Pseudomonas aeruginosa PAO1]	Pseudomonas ae...	76.3	76.3	41%	4e-13	22.63%	714	P34750.2
RecName: Full=Type IV pilus biogenesis and competence protein PilQ; Flags: Precursor [Neisseria meningitidis Z2491]	Neisseria meningi...	73.2	73.2	42%	4e-12	23.97%	761	Q9JVW4.1
RecName: Full=Type IV pilus biogenesis and competence protein PilQ; Flags: Precursor [Neisseria meningitidis MC58]	Neisseria meningi...	72.8	72.8	42%	5e-12	23.97%	769	Q70M91.2
RecName: Full=Type IV pilus biogenesis and competence protein PilQ; Flags: Precursor [Neisseria meningitidis H44/...	Neisseria meningi...	72.8	72.8	42%	5e-12	23.97%	777	Q9ZHF3.2

Feedback

Analysis of SwissProt BLAST Results:

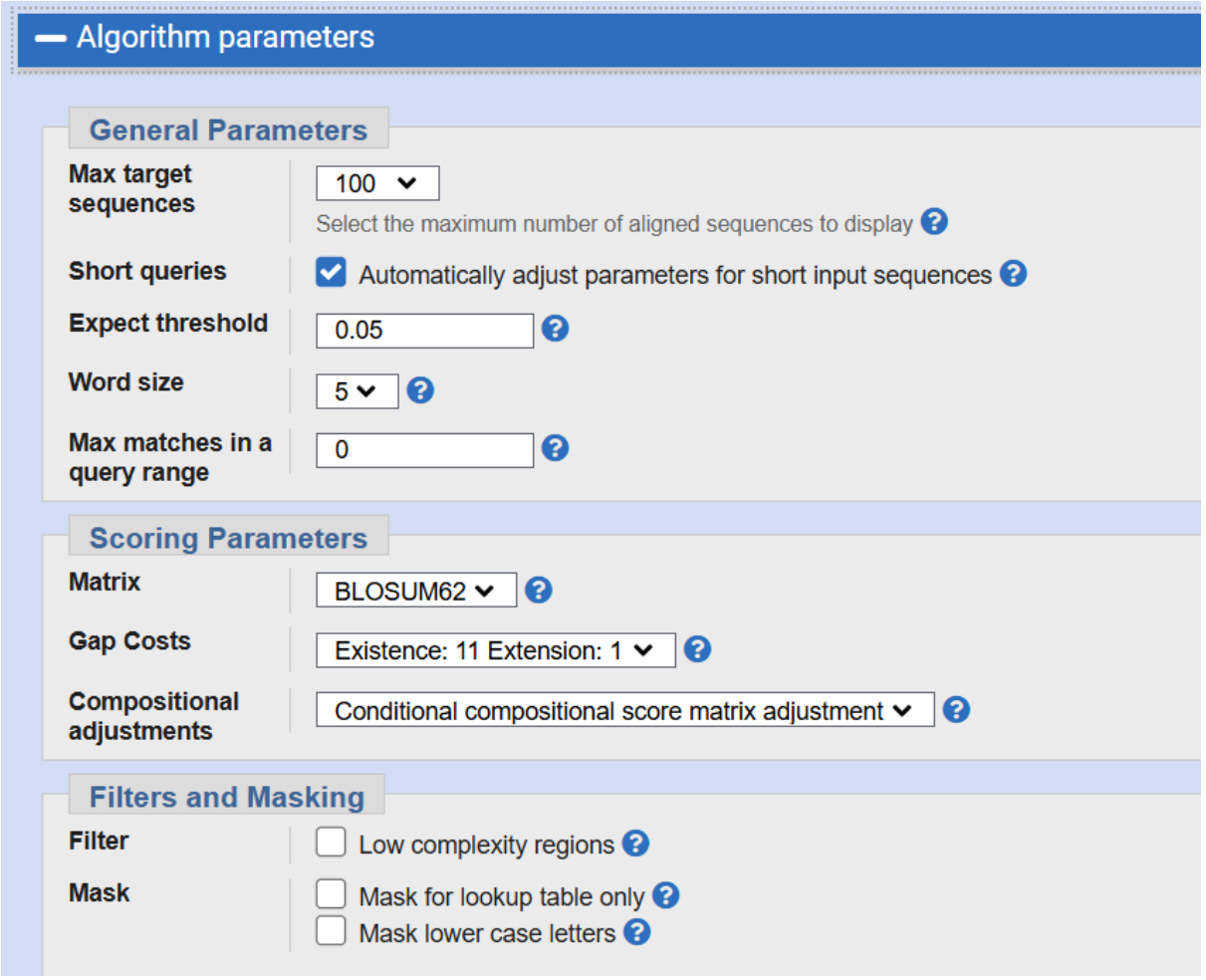
- All 33 related sequences have extremely low E values, close to zero.
- The maximum query coverage observed is 98%, associated with the Type 3 secretion system secretin from *Pseudomonas syringae pv. Syringae*.
- The minimum query coverage recorded is 23%, occurring in three sequences.
- The highest identity (44.16%) is with *Pseudomonas syringae* T3SS secretin (Q01723.2).
- Pseudomonas syringae (highest identity) is a well-known plant pathogen.
- Yersinia, Shigella, and Salmonella are human pathogens. Ralstonia and Xanthomonas are plant pathogens. Escherichia coli and Klebsiella pneumoniae are opportunistic human pathogens.
- The lowest identity percentage is 21.75%, found in the uncharacterized protein y4xJ from *Sinorhizobium fredii NGR234*.
- The low sequence identity (~30%) in some cases but functional similarity indicates potential structural conservation despite sequence divergence.

- The query protein is highly similar to T3SS secretins, particularly HrpH from *Pseudomonas syringae*.

---

2. List the algorithm parameters used for the search (Q1).

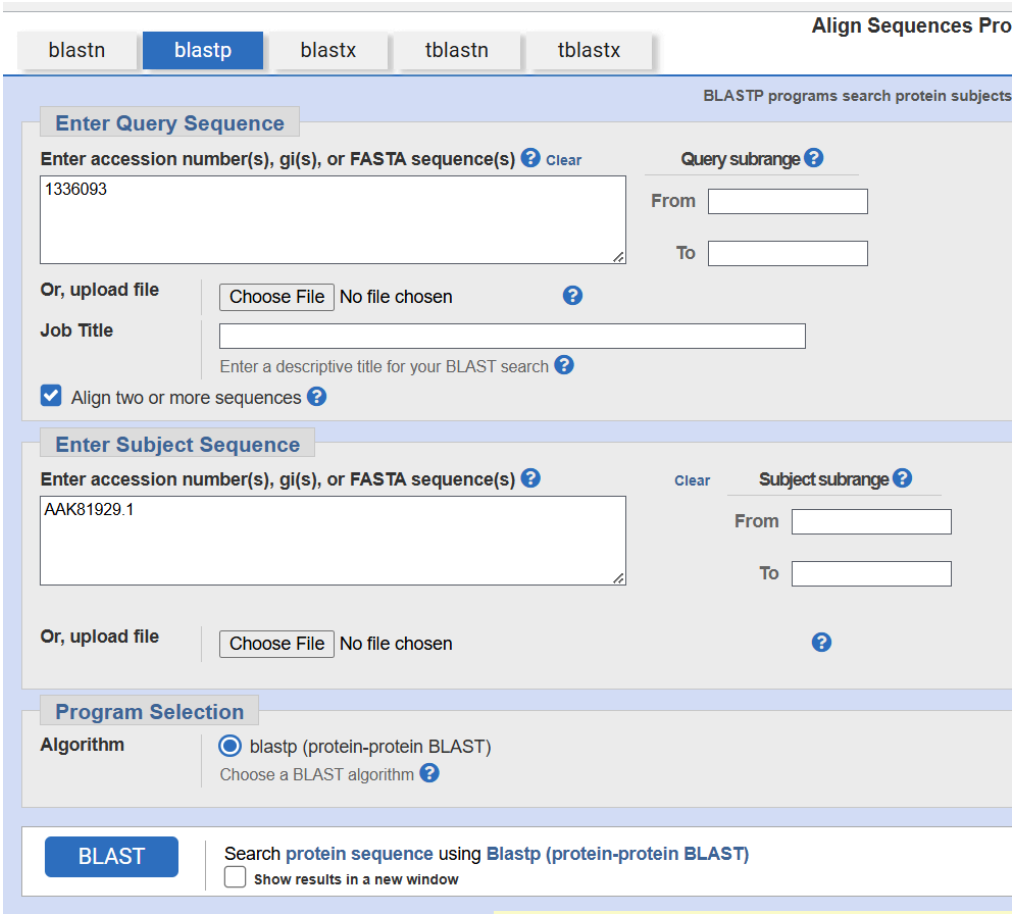
Default parameters which we used for the first question:



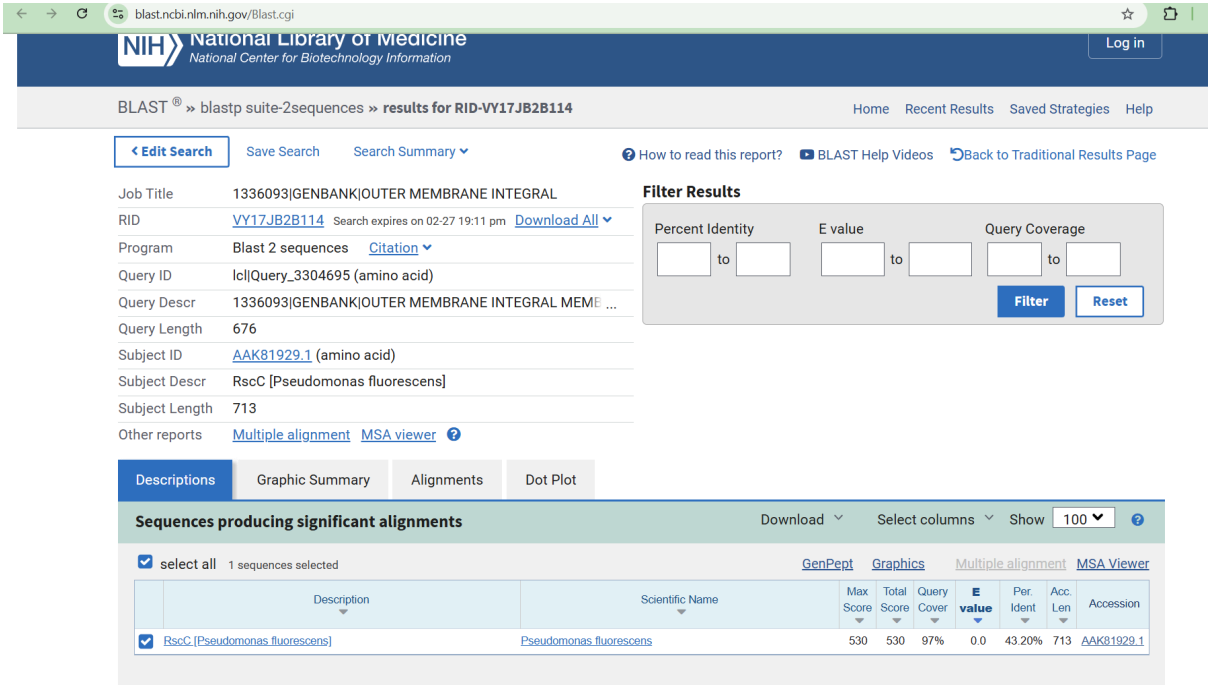
3. What is the sequence identity of the query sequence (given in Q1) with AAK81929.1?

Steps:

Use align 2 or more sequences and Run BLASTP using the HrcC protein sequence and AAK81929.1 in subject sequence.



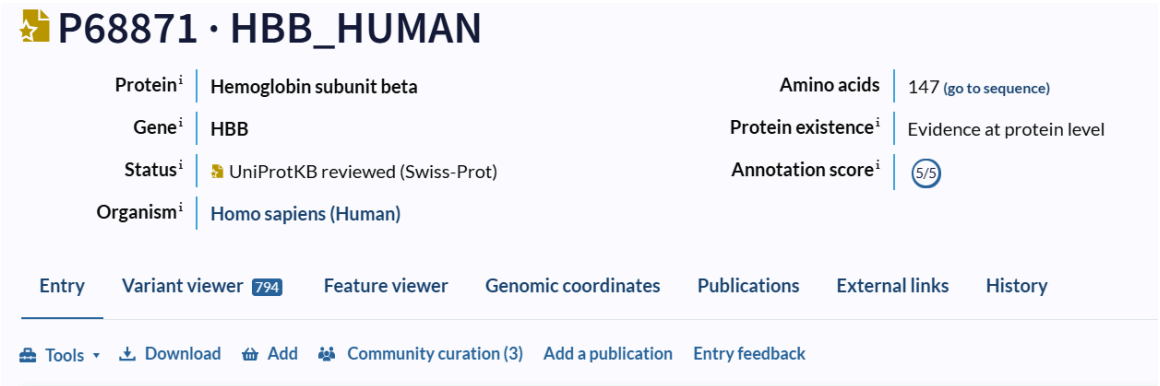
43.20% sequence identity



---

4. How far are hemoglobin (beta) sequences in humans and chicken similar?

Human hemoglobin beta: P68871.
Chicken hemoglobin beta: P02112.

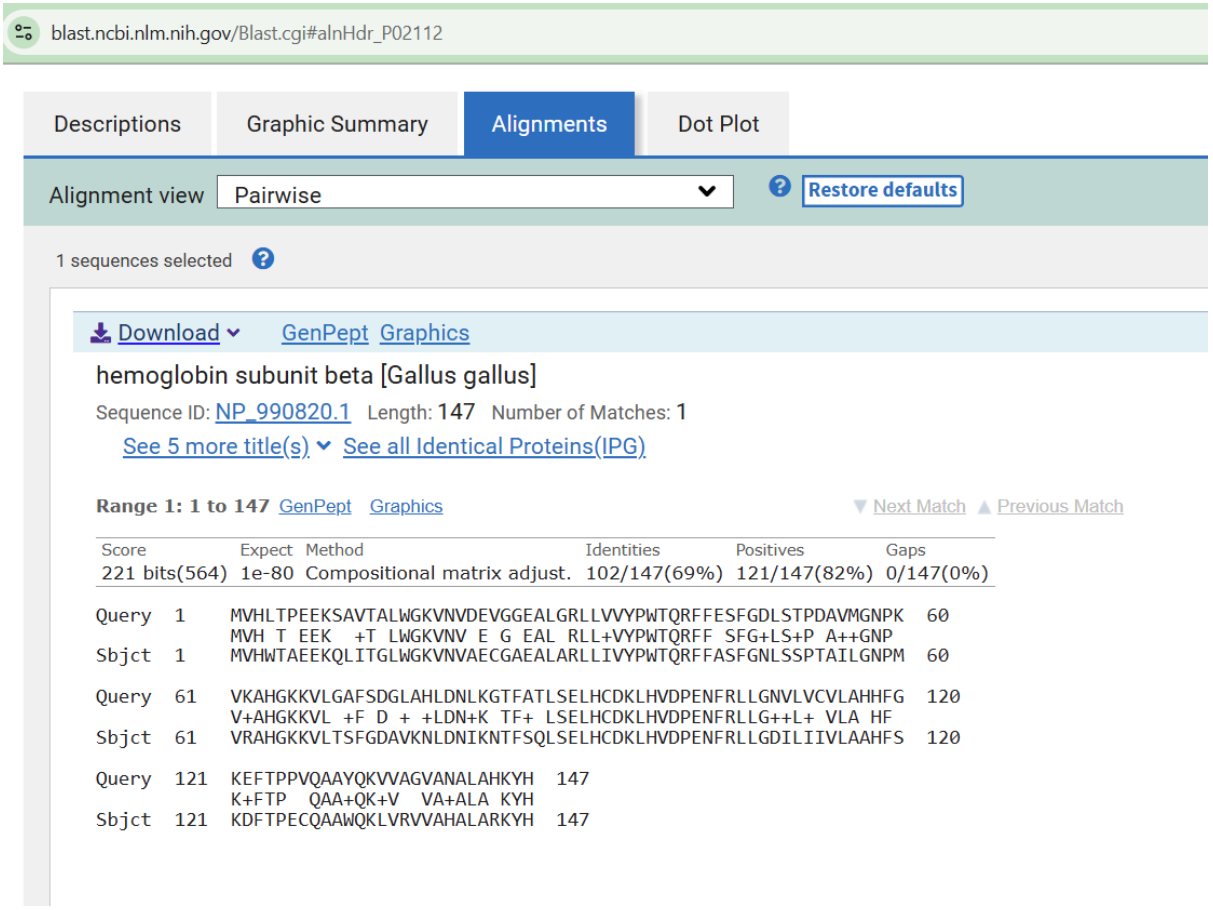


P02112 · HBB_CHICK

Protein[i]	Hemoglobin subunit beta	Amino acids	147 (go to sequence)
Gene[i]	HBB	Protein existence[i]	Evidence at protein level
Status[i]	UniProtKB reviewed (Swiss-Prot)	Annotation score[i]	5/5
Organism[i]	Gallus gallus (Chicken)		

Based on the BLASTP alignment results, the similarity between human and chicken hemoglobin beta sequences is as follows:

1. Identity: 69% (102 out of 147 amino acids are exactly the same)
2. Positives: 82% (121 out of 147 amino acids share similar chemical properties, meaning functional similarity)
3. Gaps: 0% (No insertions or deletions, indicating a well-aligned sequence)
4. E-value: 1e-80 (Highly significant match, indicating evolutionary conservation)



---

5. Write a program to list all the matching pentapeptides (which occur in both the sequences) and their frequency of occurrence in given sequences.
Human hemoglobin beta: P68871.
Chicken hemoglobin beta: P02112.

<u>Got these protein sequence from uniprot:</u>

>sp|P02112|HBB_CHICK Hemoglobin subunit beta OS=Gallus gallus OX=9031 GN=HBB PE=1 SV=2
MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM
VRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS
KDFTPECQAAWQKLVRVVAHALARKYH

>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens OX=9606 GN=HBB PE=1 SV=2

MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH

This code identifies common pentapeptides (five-amino-acid sequences) between human and chicken hemoglobin sequences. It does this by generating all possible pentapeptides from both sequences and counting their occurrences using Python's Counter from the collections module.

The program then finds overlapping pentapeptides and displays their frequencies in each sequence. This analysis helps compare structural similarities between the two proteins, revealing conserved regions that might be functionally important.

p5_BS22B009.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text

```python
#Question5
from collections import Counter

def get_pentapeptide_counts(sequence):
    pentapeptides = [sequence[i:i+5] for i in range(len(sequence) - 4)]
    return Counter(pentapeptides)

human_seq = "MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK"
human_seq += "VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG"
human_seq += "KEFTPPVQAAYQKVVAGVANALAHKYH"

chicken_seq = "MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM"
chicken_seq += "VRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS"
chicken_seq += "KDFTPECQAAWQKLVRVVAHALARKYH"

human_counts = get_pentapeptide_counts(human_seq)
chicken_counts = get_pentapeptide_counts(chicken_seq)

common_pentapeptides = set(human_counts.keys()) & set(chicken_counts.keys())

print("Matching Pentapeptides and Their Frequency:")
for peptide in sorted(common_pentapeptides):
    human_count = human_counts[peptide]
    chicken_count = chicken_counts[peptide]
    total_count = human_count + chicken_count
    print(f"{peptide} occurs {total_count} time(s) in both peptides "
          f"({human_count} time(s) in Seq. 1 and {chicken_count} time(s) in Seq. 2)")
```

Result for the code:

```
Matching Pentapeptides and Their Frequency:
AHGKK occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
CDKLH occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
DKLHV occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
DPENF occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
ELHCD occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
ENFRL occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
FRLLG occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
GKKVL occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
GKVNV occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
HCDKL occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
HGKKV occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
HVDPE occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
KLHVD occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
LHCDK occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
LHVDP occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
LSELH occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
LWGKV occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
NFRLL occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
PENFR occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
PWTQR occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
SELHC occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
TQRFF occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
VDPEN occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
VYPWT occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
WGKVN occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
WTQRF occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
YPWTQ occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
```

---

6. Write a program to compute sequence identity, similarity, query coverage and gap percentage from the alignment of human and chicken hemoglobin sequences (refer Q4).

Here's my Python program to compute sequence identity, similarity, query coverage, and gap percentage between human and chicken hemoglobin sequences. I used a character-by-character comparison, considering exact matches for identity and grouping similar amino acids for similarity. Query coverage is calculated based on the aligned portion, and gaps are accounted for separately.

It calculates four key metrics:

1. Sequence Identity – The percentage of exact matches between the two sequences.
2. Sequence Similarity – The percentage of residues that either match exactly or belong to the same biochemical similarity group.

3. Query Coverage – The proportion of the query sequence (human hemoglobin) that aligns with the target sequence (chicken hemoglobin).
4. Gap Percentage – The fraction of positions where either sequence has a gap ('-').

To determine similarity, I grouped amino acids based on their properties, such as hydrophobicity and charge. Then, I iterated through the sequences to count matches, similar residues, and gaps. Finally, I calculated and printed the results.

p5_BS22B009.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text

```python
human_seq = (
    "MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK"
    "VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG"
    "KEFTPPVQAAYQKVVAGVANALAHKYH"
)

chicken_seq = (
    "MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM"
    "VRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS"
    "KDFTPECQAAWQKLVRVVAHALARKYH"
)

def calculate_metrics(seq1, seq2):
    matches = 0
    similar = 0
    gaps = 0
    total_length = min(len(seq1), len(seq2))
    similarity_groups = [
        set("NDEQ"), set("MILV"), set("FYW"), set("KRH"), set("ST"), set("AGP")
    ]
    for i in range(total_length):
        if seq1[i] == seq2[i]:
            matches += 1
            similar += 1
        elif seq1[i] == '-' or seq2[i] == '-':
            gaps += 1
        else:
            for group in similarity_groups:
                if seq1[i] in group and seq2[i] in group:
                    similar += 1
                    break
    identity = (matches / total_length) * 100
    similarity = (similar / total_length) * 100
    query_coverage = (total_length / len(seq1)) * 100
    gap_percentage = (gaps / total_length) * 100
```

```python
    identity = (matches / total_length) * 100
    similarity = (similar / total_length) * 100
    query_coverage = (total_length / len(seq1)) * 100
    gap_percentage = (gaps / total_length) * 100
    return identity, similarity, query_coverage, gap_percentage

identity, similarity, query_coverage, gap_percentage = calculate_metrics(human_seq, chicken_seq)

print(f"Sequence Identity: {identity:.2f}%")
print(f"Sequence Similarity: {similarity:.2f}%")
print(f"Query Coverage: {query_coverage:.2f}%")
print(f"Gap Percentage: {gap_percentage:.2f}%")
```

Result:

```
Sequence Identity: 69.39%
Sequence Similarity: 83.67%
Query Coverage: 100.00%
Gap Percentage: 0.00%
```

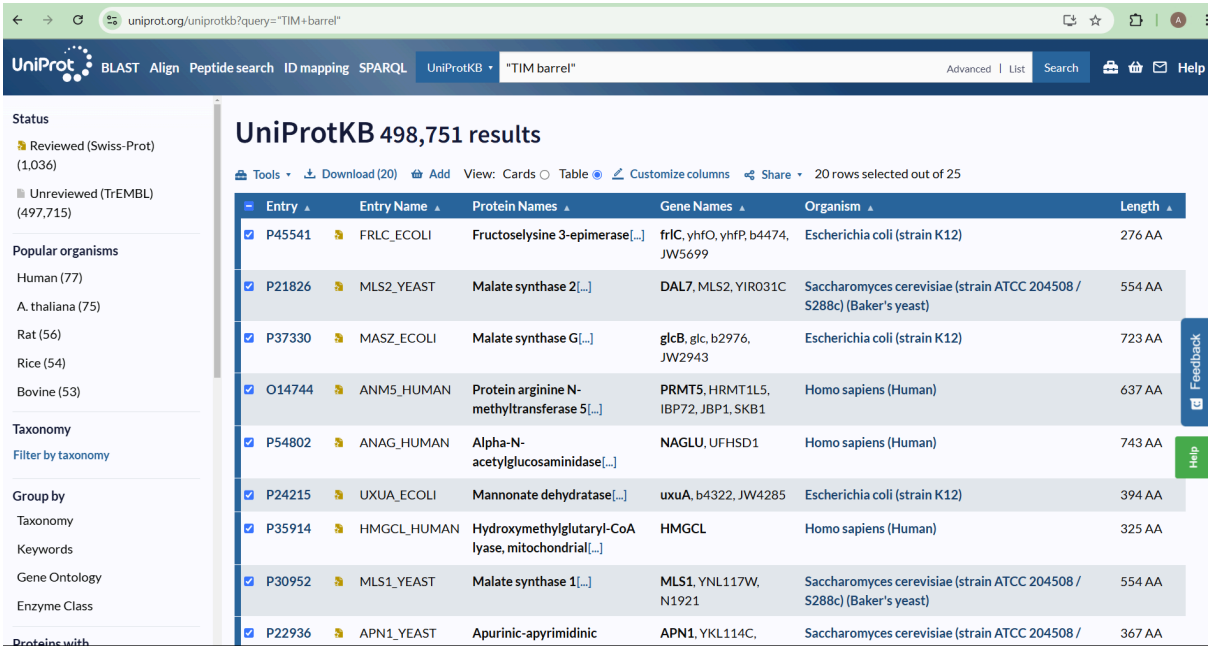
---

7. Obtain the multiple sequence alignment for TIM barrel proteins from different organisms (select 20 proteins, for example). Compare the results obtained with Clustal Omega, MAFFT, and MUSCLE. List 5 residue positions which are aligned differently in these three methods.

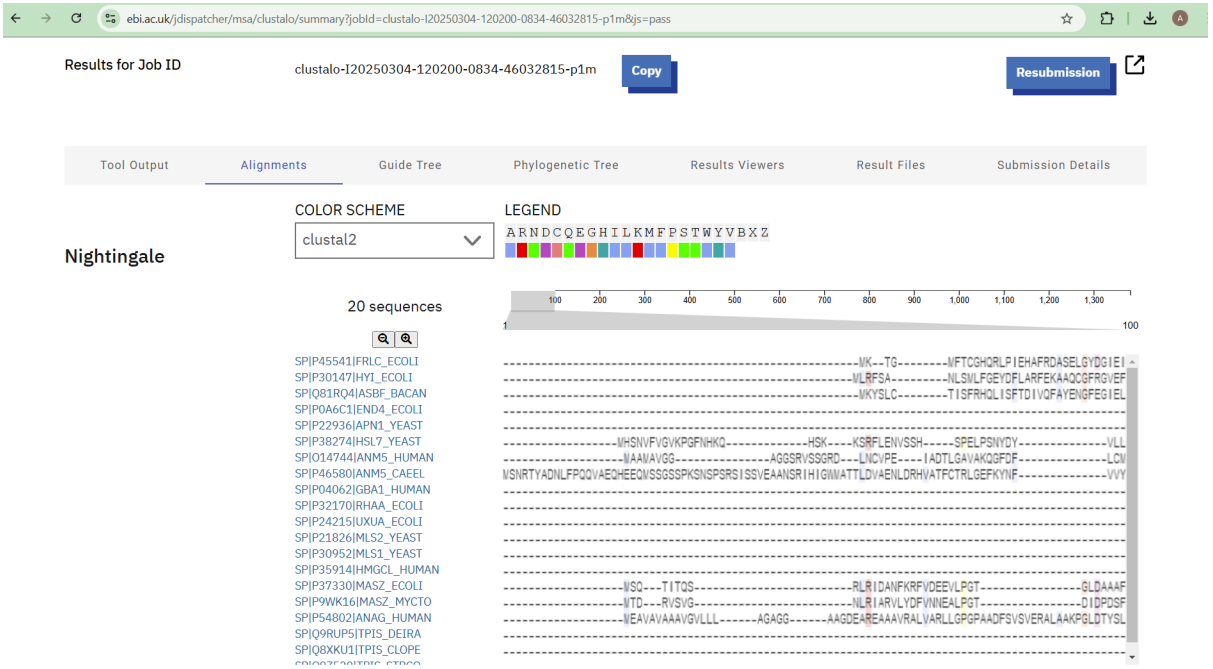
<u>Steps:</u>

1.Collect 20 TIM barrel protein sequences from UniProt.
2. Use three alignment tools:
   - Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/)
   - MAFFT (https://mafft.cbrc.jp/alignment/server/)
   - MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/)
3. Compare alignments and list 5 residue positions with differences.

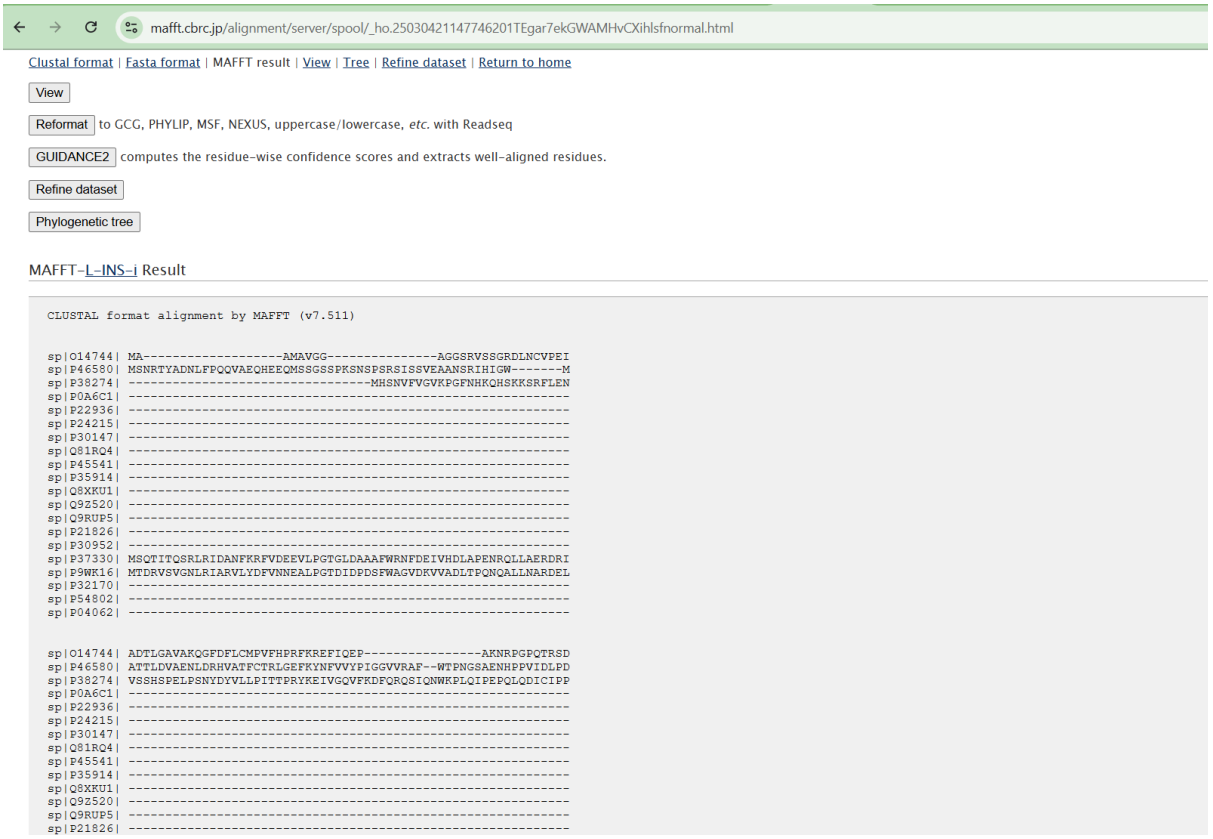
20 sequences download in FASTA format from Uniprot :



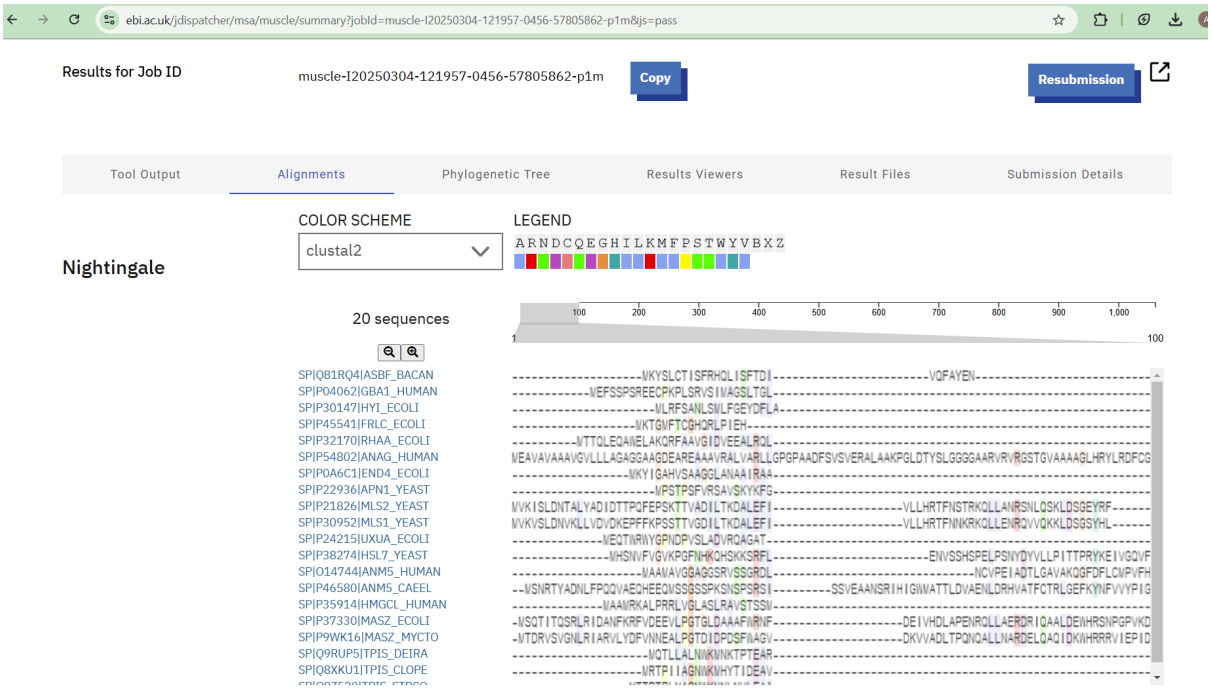
Multiple sequence alignment (MSA) with Clustal Omega.



Multiple sequence alignment (MSA) with MAFFT:



Multiple sequence alignment (MSA) with MUSCLE:



5 residue positions that are aligned differently in these three methods are 1, 2, 3, 4, 5.

1. Residue Position 1:

    - Clustal Omega aligns M (Methionine) in P46580, while the other sequences have gaps.
    - MAFFT aligns M in Q9Z520, with gaps in other sequences.
    - MUSCLE aligns M in P54802, while others have gaps.
2. Residue Position 2:

    - Clustal Omega places S (Serine) in P46580, with gaps in the other sequences.
    - MAFFT aligns T (Threonine) in Q9Z520, with gaps in the rest.
    - MUSCLE aligns E (Glutamic Acid) in P54802, while others have gaps.
3. Residue Position 3:

    - Clustal Omega aligns N (Asparagine) in P46580, with gaps elsewhere.
    - MAFFT assigns T (Threonine) in Q9Z520, while M (Methionine) is present in other sequences.
    - MUSCLE places A (Alanine) in P54802, with gaps in the remaining sequences.
4. Residue Position 4:

    - Clustal Omega aligns R (Arginine) in P46580, with gaps in the other sequences.
    - MAFFT shows different amino acids across different sequences, meaning no consistent alignment.
    - MUSCLE aligns V (Valine) in P54802, while others have gaps.
5. Residue Position 5:

    - Clustal Omega aligns T (Threonine) in P46580, with gaps in the other sequences.
    - MAFFT presents different amino acids across different sequences, showing variation.
    - MUSCLE places A (Alanine) in P54802, with gaps in the rest.

Conclusion:

- Clustal Omega, MAFFT, and MUSCLE show significant differences in alignment at these five positions.

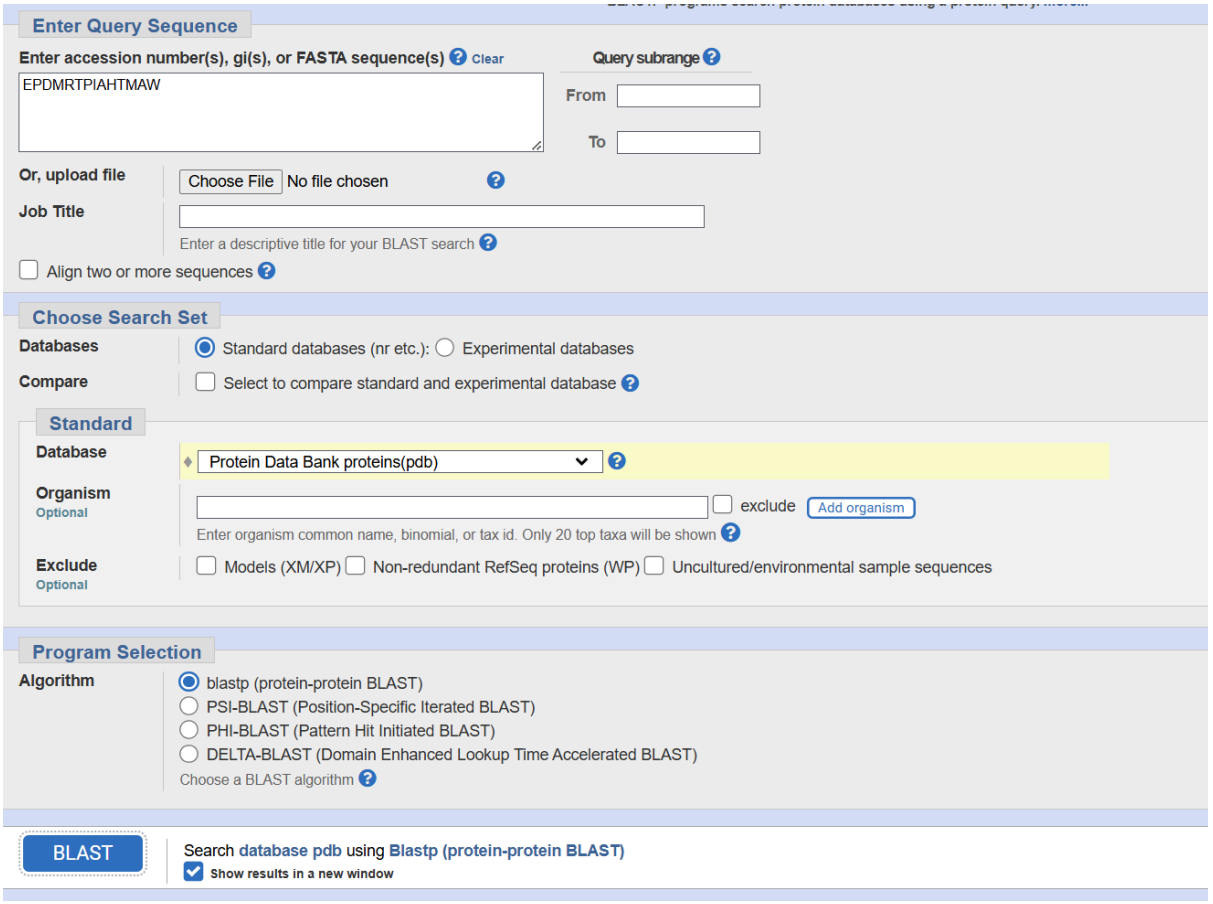
- Clustal Omega and MUSCLE tend to align specific residues while keeping gaps in other sequences, whereas MAFFT exhibits more variation in residue choices.
- MAFFT shows the highest variability, while Clustal Omega and MUSCLE align specific amino acids with more gaps in others.
- These differences highlight the varying sensitivity and algorithmic approaches of multiple sequence alignment (MSA) tools.

---

8. Blast the below sequence 'EPDMRTPIAHTMAW' against the PDB database. Analyze the results and discuss the significance of the results.

<u>Steps:</u>

Go to NCBI BLAST → Select BLASTP.
2. Enter sequence 'EPDMRTPIAHTMAW'.
3. Select PDB (Protein Data Bank) database.
4. Run the search and analyze:
  - Top hits
  - E-value, identity, and coverage



Blast results:

blast.ncbi.nlm.nih.gov/Blast.cgi

Other reports Distance tree of results Multiple alignment MSA viewer

Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all 100 sequences selected GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Escherichia coli str. K-12 substr. W3110]	Escherichia coli s...	53.2	53.2	100%	2e-10	100.00%	424	2EGH_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Escherichia coli K-12]	Escherichia coli ...	53.2	53.2	100%	2e-10	100.00%	420	3ANL_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Escherichia coli K-12]	Escherichia coli ...	53.2	53.2	100%	2e-10	100.00%	410	3R0I_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Escherichia coli]	Escherichia coli	53.2	53.2	100%	2e-10	100.00%	406	1Q0L_A
Chain A, 1-deoxy-D-xylulose-5-phosphate reductoisomerase [Escherichia coli K-12]	Escherichia coli ...	53.2	53.2	100%	2e-10	100.00%	398	1K5H_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Escherichia coli]	Escherichia coli	53.2	53.2	100%	2e-10	100.00%	398	1T1R_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Escherichia coli]	Escherichia coli	43.1	43.1	100%	9e-07	85.71%	406	1Q0H_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Escherichia coli]	Escherichia coli	43.1	43.1	100%	9e-07	85.71%	400	1JVS_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Zymomonas mobilis]	Zymomonas mob...	42.2	42.2	93%	2e-06	84.62%	388	1R0K_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Yersinia pseudotuberculosis YPIII]	Yersinia pseudot...	40.9	40.9	86%	5e-06	91.67%	401	3IIE_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Acinetobacter baumannii AB307-0294]	Acinetobacter ba...	35.4	35.4	93%	4e-04	76.92%	406	4ZN6_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Moraxella catarrhalis]	Moraxella catarrh...	33.3	33.3	86%	0.003	75.00%	432	4ZQE_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Moraxella catarrhalis]	Moraxella catarrh...	33.3	33.3	86%	0.003	75.00%	415	4ZQG_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Vibrio vulnificus CMCP6]	Vibrio vulnificus ...	28.2	28.2	86%	0.17	75.00%	427	5KQO_A
Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Vibrio vulnificus CMCP6]	Vibrio vulnificus ...	28.2	28.2	86%	0.17	75.00%	405	5KRR_A
Chain A, Glutamate dehydrogenase 2 [Arabidopsis thaliana]	Arabidopsis thali...	26.1	39.0	93%	0.95	69.23%	414	8OWM_A
Chain A, 1-deoxy-D-xylulose-5-phosphate reductoisomerase [Toxoplasma gondii]	Toxoplasma gondii	24.8	24.8	93%	2.7	53.85%	472	8S65_A
Chain A, Acyl-CoA dehydrogenase [Streptomyces ficellus]	Streptomyces fic...	24.4	24.4	43%	3.8	100.00%	384	8HK0_A
Chain A, Glutamate dehydrogenase 1 [Arabidopsis thaliana]	Arabidopsis thali...	24.4	37.3	93%	3.8	64.29%	414	6YEH_A
Chain Bh, 30S ribosomal protein S9, mitochondrial [Arabidopsis thaliana]	Arabidopsis thali...	24.4	24.4	43%	3.8	100.00%	430	6XYW_Bh

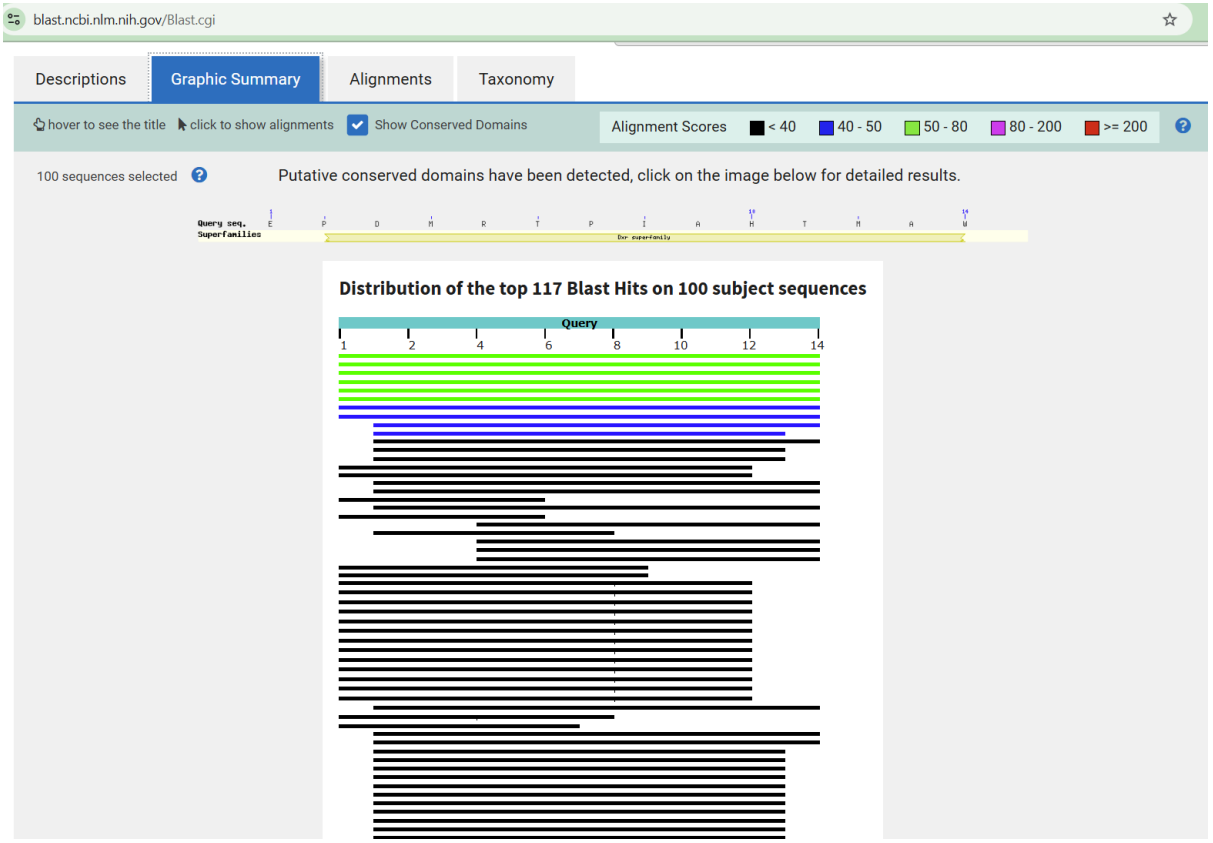
The BLAST results indicate that the query sequence "EPDMRTPIAHTMAW" has 100% query coverage and high sequence identity with 1-deoxy-D-xylulose 5-phosphate reductoisomerase (DXR) in *Escherichia coli* and other bacterial species.

Sequences with lower scores, such as *Zymomonas mobilis*, *Yersinia pseudotuberculosis*, and *Moraxella catarrhalis*, still show high sequence identity (ranging from 85.71% to 100%), indicating a strong conservation of this sequence across different bacterial species.

The Query Cover is 100% for the top sequences, which means that the entire query sequence (all 14 amino acids) aligns completely with the matched sequences in the database.

Some of the lower-ranked hits (e.g., *Arabidopsis thaliana*, *Toxoplasma gondii*) show slightly lower query coverage (~93% or lower), meaning only part of the sequence aligns with their respective proteins.

The E-value (Expect value) for the top matches is 2e-10, which is extremely low. This indicates a highly significant match, meaning the alignment is not due to random chance.



The graphic summary shows that the sequence aligns well with members of the DXR superfamily, highlighting its evolutionary conservation. Given its high sequence identity and conservation, this sequence could be essential for enzyme function or stability.

This suggests the query sequence is part of a highly conserved functional motif in DXR, a key enzyme in the non-mevalonate pathway of isoprenoid biosynthesis. Identifying such a conserved motif with just a short sequence demonstrates the power of sequence alignment in detecting functionally related proteins, which could be useful for drug targeting in bacterial infections.