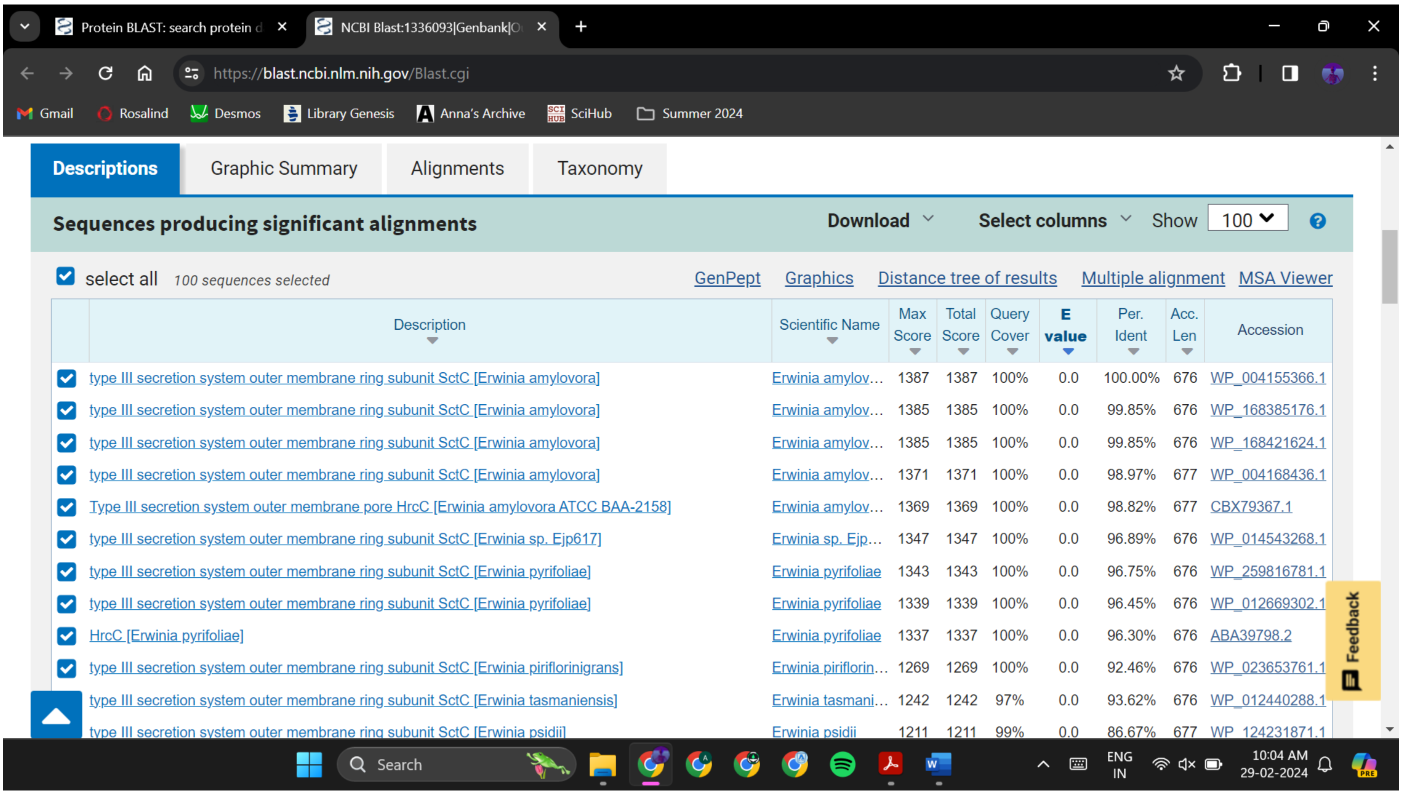
Anirudh Rao
BE21B004

BT3040 – Bioinformatics

Practical 5

1

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



Sequences producing significant alignments

☑ select all *100 sequences selected*

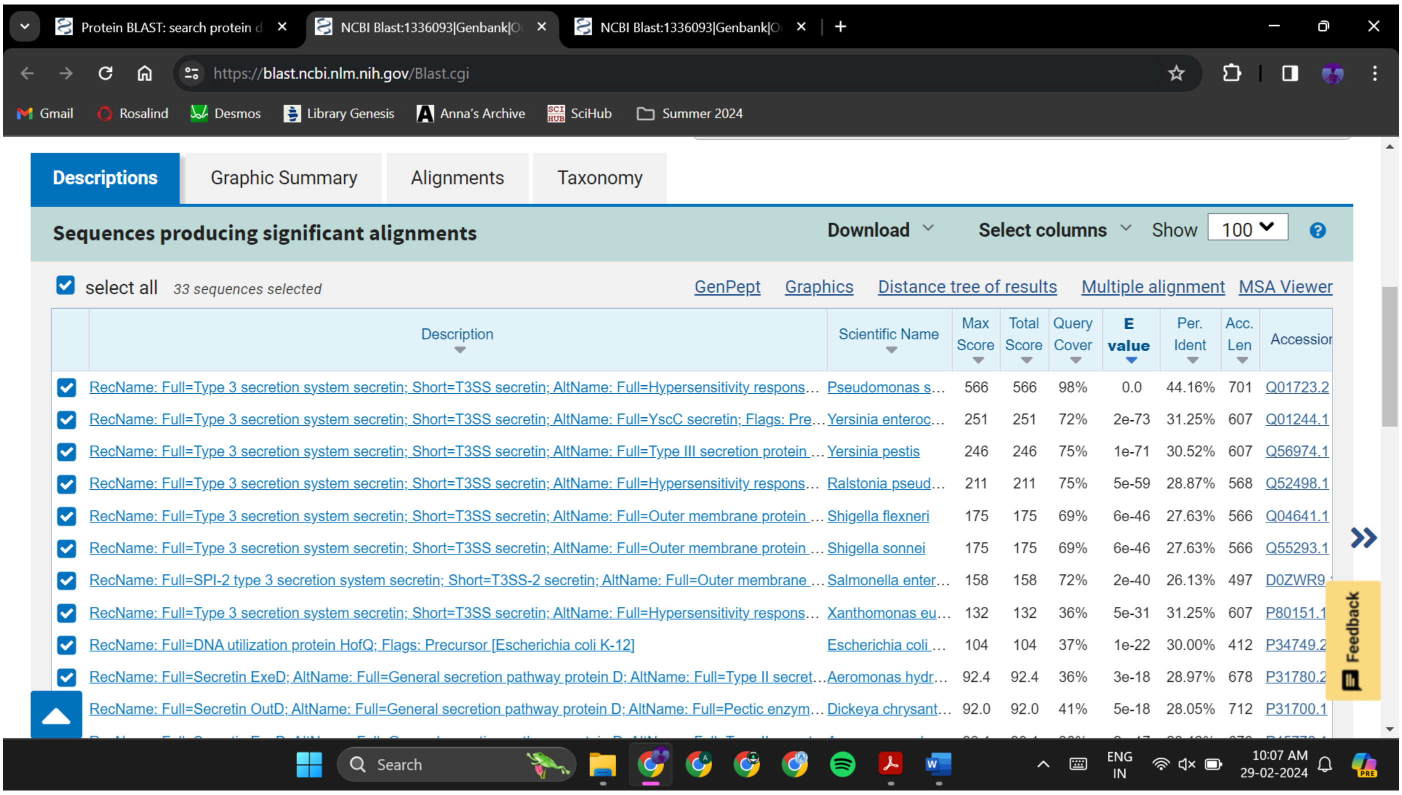
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]	Erwinia amylov...	1387	1387	100%	0.0	100.00%	676	WP_004155366.1
☑	type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]	Erwinia amylov...	1385	1385	100%	0.0	99.85%	676	WP_168385176.1
☑	type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]	Erwinia amylov...	1385	1385	100%	0.0	99.85%	676	WP_168421624.1
☑	type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]	Erwinia amylov...	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 amylov...	1369	1369	100%	0.0	98.82%	677	CBX79367.1
☑	type III secretion system outer membrane ring subunit SctC [Erwinia sp. Ejp617]	Erwinia sp. Ejp...	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 piriflorin...	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	99%	0.0	86.67%	677	WP_124231871.1

Analysis

- All 100 similar sequences have an E value of 0
- 29 sequences have 100% query coverage
- The lowest query coverage is 90%
- Only 1 sequence has 100% identity (type III secretion system outer membrane ring subunit SctC [Erwinia amylovora])
- Lowest percentage identity is 66.96% (type III secretion system outer membrane ring subunit SctC [Dickeya lacustris])

Thus, the given sequence is likely to be type III secretion system outer membrane ring subunit SctC [Erwinia amylovora].

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



Sequences producing significant alignments

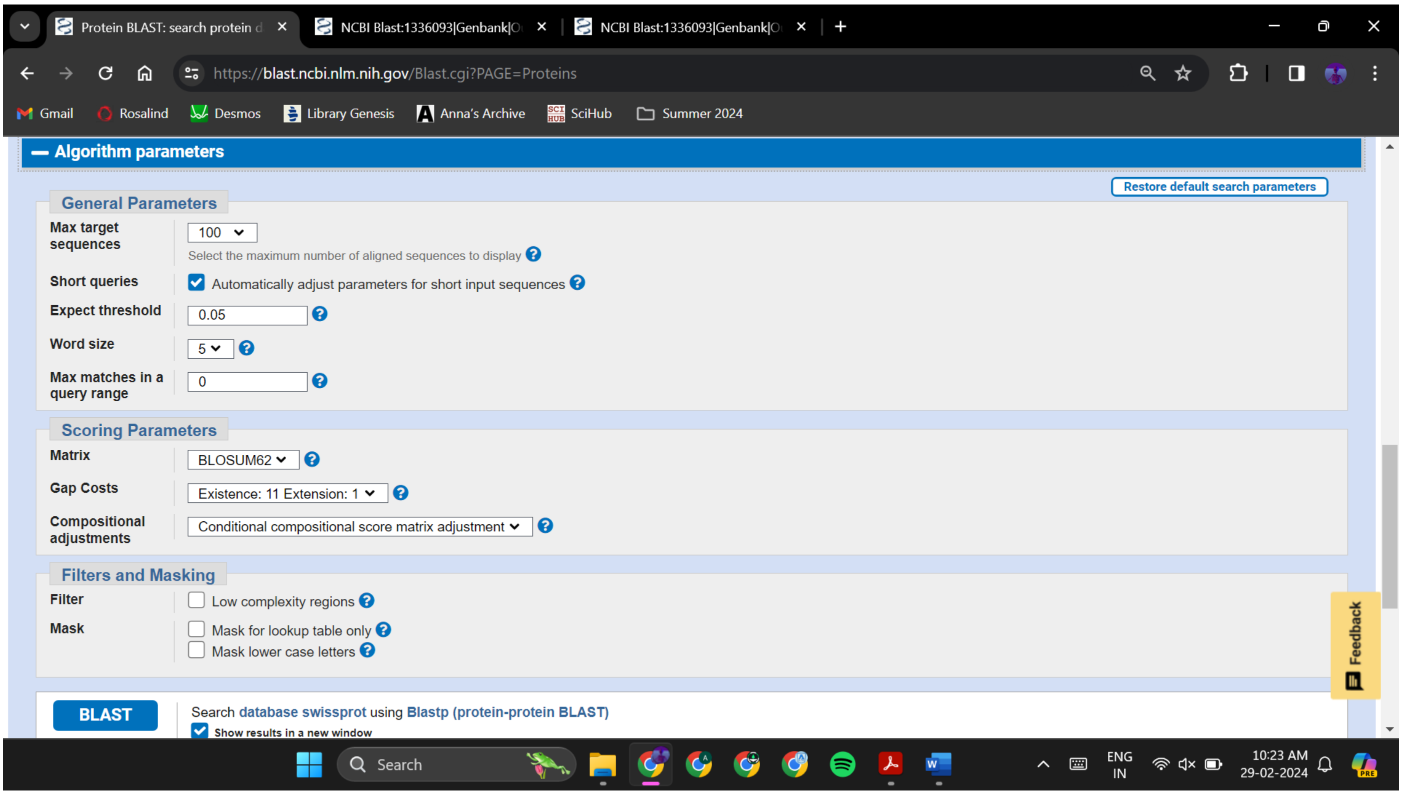
☑ select all *33 sequences selected*

	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 respons...	Pseudomonas s...	566	566	98%	0.0	44.16%	701	Q01723.2
☑	RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=YscC secretin; Flags: Pre...	Yersinia enteroc...	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 ...	Yersinia pestis	246	246	75%	1e-71	30.52%	607	Q56974.1
☑	RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Hypersensitivity respons...	Ralstonia pseud...	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 ...	Shigella flexneri	175	175	69%	6e-46	27.63%	566	Q04641.1
☑	RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Outer membrane protein ...	Shigella sonnei	175	175	69%	6e-46	27.63%	566	Q55293.1
☑	RecName: Full=SPI-2 type 3 secretion system secretin; Short=T3SS-2 secretin; AltName: Full=Outer membrane ...	Salmonella enter...	158	158	72%	2e-40	26.13%	497	D0ZWR9.
☑	RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Hypersensitivity respons...	Xanthomonas eu...	132	132	36%	5e-31	31.25%	607	P80151.1
☑	RecName: Full=DNA utilization protein HofQ; Flags: Precursor [Escherichia coli K-12]	Escherichia coli ...	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 secret...	Aeromonas hydr...	92.4	92.4	36%	3e-18	28.97%	678	P31780.2
	RecName: Full=Secretin OutD; AltName: Full=General secretion pathway protein D; AltName: Full=Pectic enzym...	Dickeya chrysant...	92.0	92.0	41%	5e-18	28.05%	712	P31700.1

Analysis

- All 33 similar sequences have E values close to 0
- The highest query coverage is 98% (Type 3 secretion system secretin [Pseudomonas syringae pv. syringae])
- The lowest query coverage is 23%, which is the coverage for 3 sequences
- The highest percentage identity is 44.16% (Type 3 secretion system secretin [Pseudomonas syringae pv. syringae])
- The lowest percentage identity is 21.75% (Uncharacterized protein y4xJ [Sinorhizobium fredii NGR234])

The algorithm parameters used in the previous question are shown below:


Protein BLAST: search protein d
NCBI Blast:1336093|Genbank|O
NCBI Blast:1336093|Genbank|O
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
Gmail
Rosalind
Desmos
Library Genesis
Anna's Archive
SciHub
Summer 2024
Restore default search parameters
Feedback
Search
ENG
IN
10:23 AM
29-02-2024

— 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: 0.05

Word size: 5

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

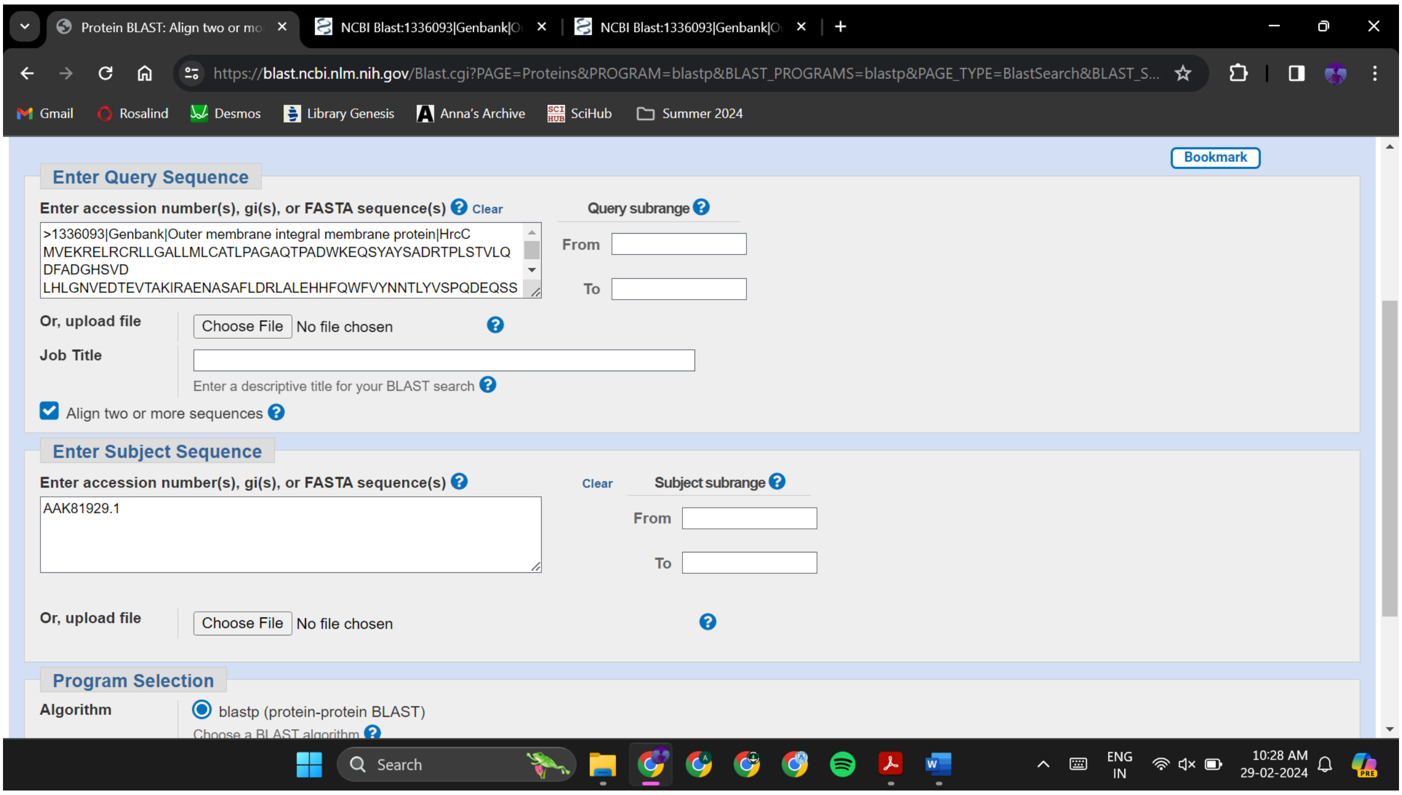
Filters and Masking

Filter: □ Low complexity regions

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

BLAST Search database swissprot using Blastp (protein-protein BLAST)
☑ Show results in a new window

The sequency identity of the query sequence with AAK81929.1 was found by selecting the “Align two or more sequences” option in BLASTP:


Protein BLAST: Align two or mo
NCBI Blast:1336093|Genbank|O
NCBI Blast:1336093|Genbank|O
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&BLAST_S...
Gmail
Rosalind
Desmos
Library Genesis
Anna's Archive
SciHub
Summer 2024
Bookmark
Search
ENG
IN
10:28 AM
29-02-2024

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>1336093|Genbank|Outer membrane integral membrane protein|HrcC
MVEKRELRCRLLGALLMLCATLPAGAQTPADWKEQSYAYSADRTPLSTVLQ
DFADGHSVD
LHLGNVEDTEVTAKIRAENASAFLDRLALEHHFQWFVYNNTLYVSPQDEQSS
```

Query subrange
From
To

Or, upload file: Choose File No file chosen

Job Title
Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

AAK81929.1

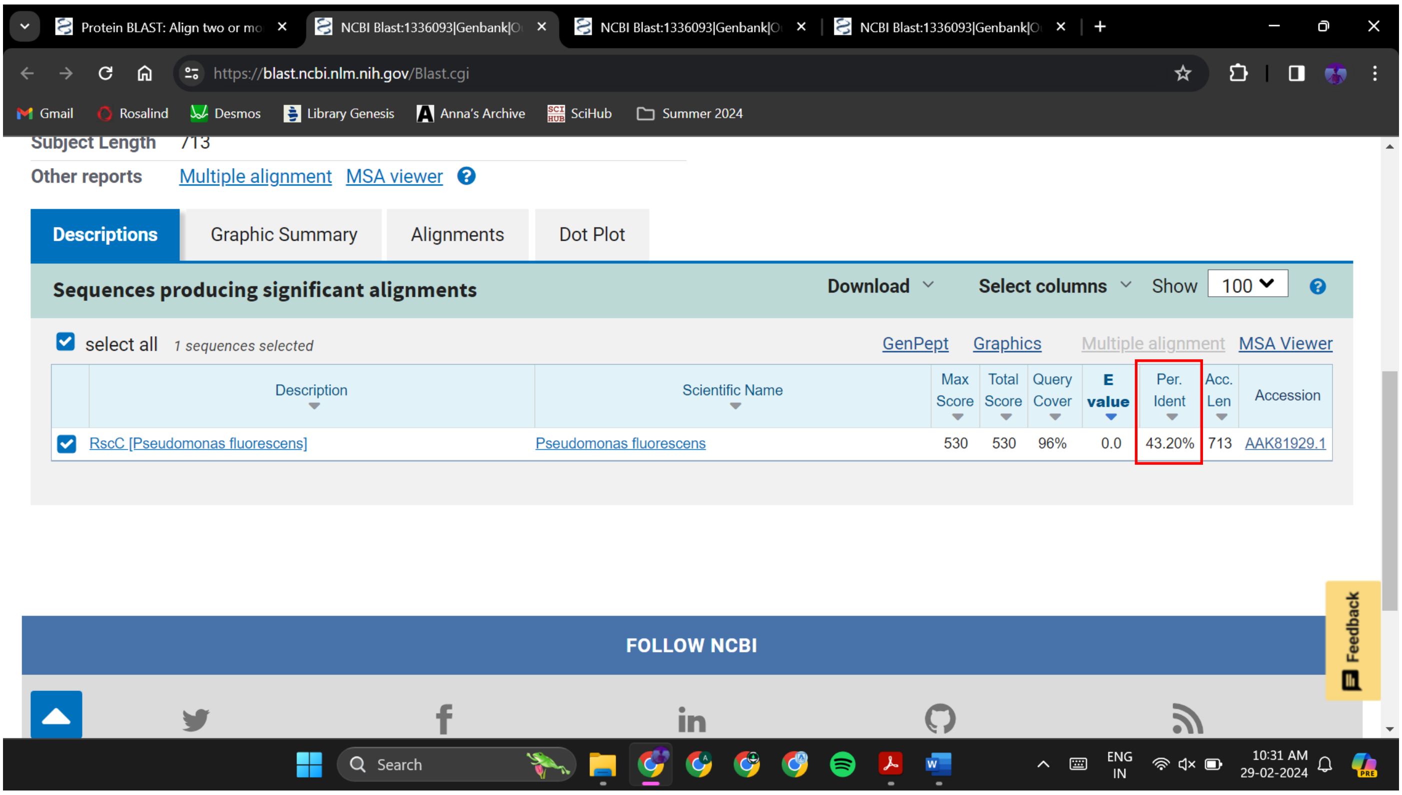
Subject subrange
From
To

Or, upload file: Choose File No file chosen

Program Selection

Algorithm: ● blastp (protein-protein BLAST)
Choose a BLAST algorithm

The percentage identity was found to be 43.20%.



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

Subject Length 713

Other reports Multiple alignment MSA viewer

Descriptions | Graphic Summary | Alignments | Dot Plot

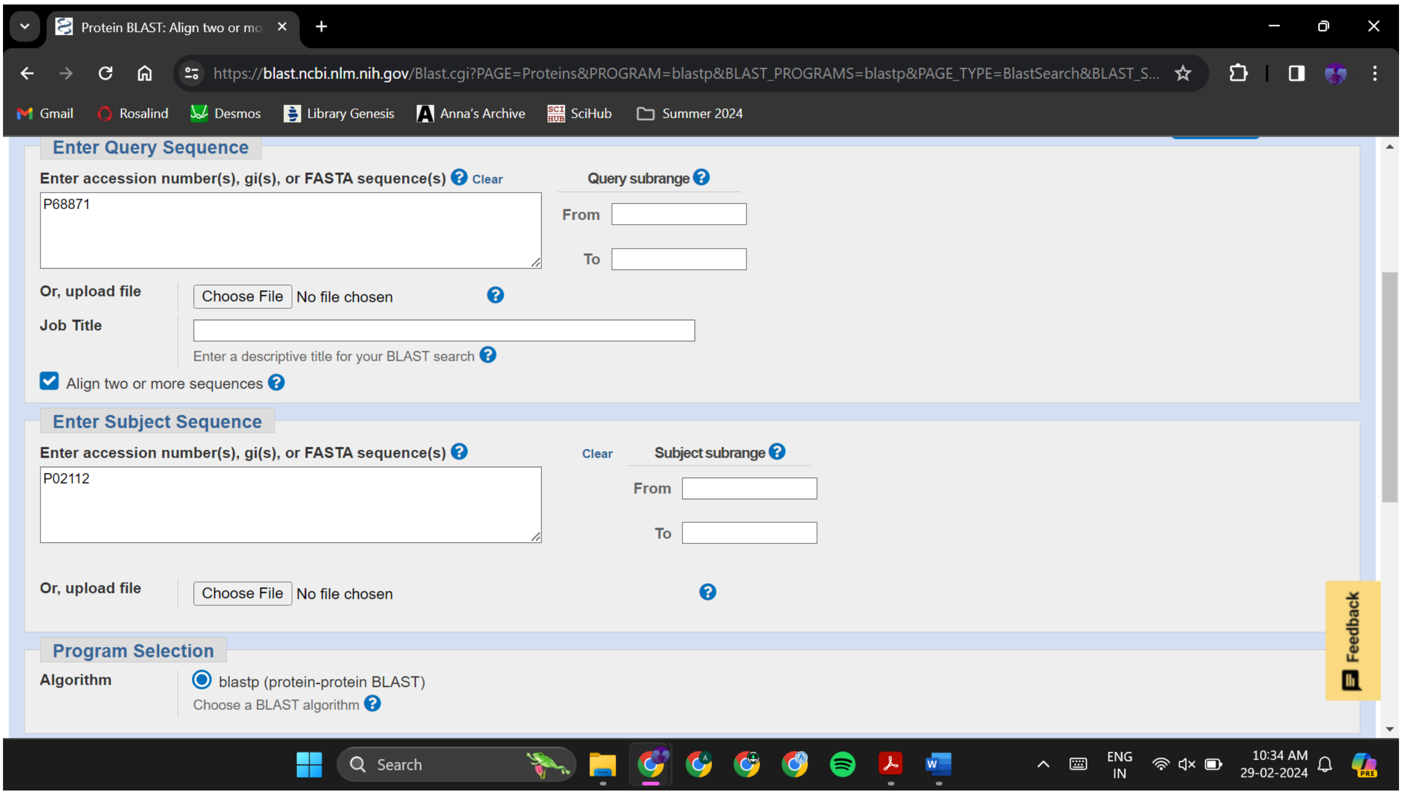
Sequences producing significant alignments

☑ select all 1 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	RscC [Pseudomonas fluorescens]	Pseudomonas fluorescens	530	530	96%	0.0	43.20%	713	AAK81929.1

Using UniProtKB, the accession number of human hemoglobin beta was found to be P68871, while that of chicken hemoglobin beta was found to be P02112.

These accession numbers were used along with the “Align two or more sequences” option in BLASTP:



https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&BLAST_S...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

P68871

Query subrange From To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

P02112

Subject subrange From To

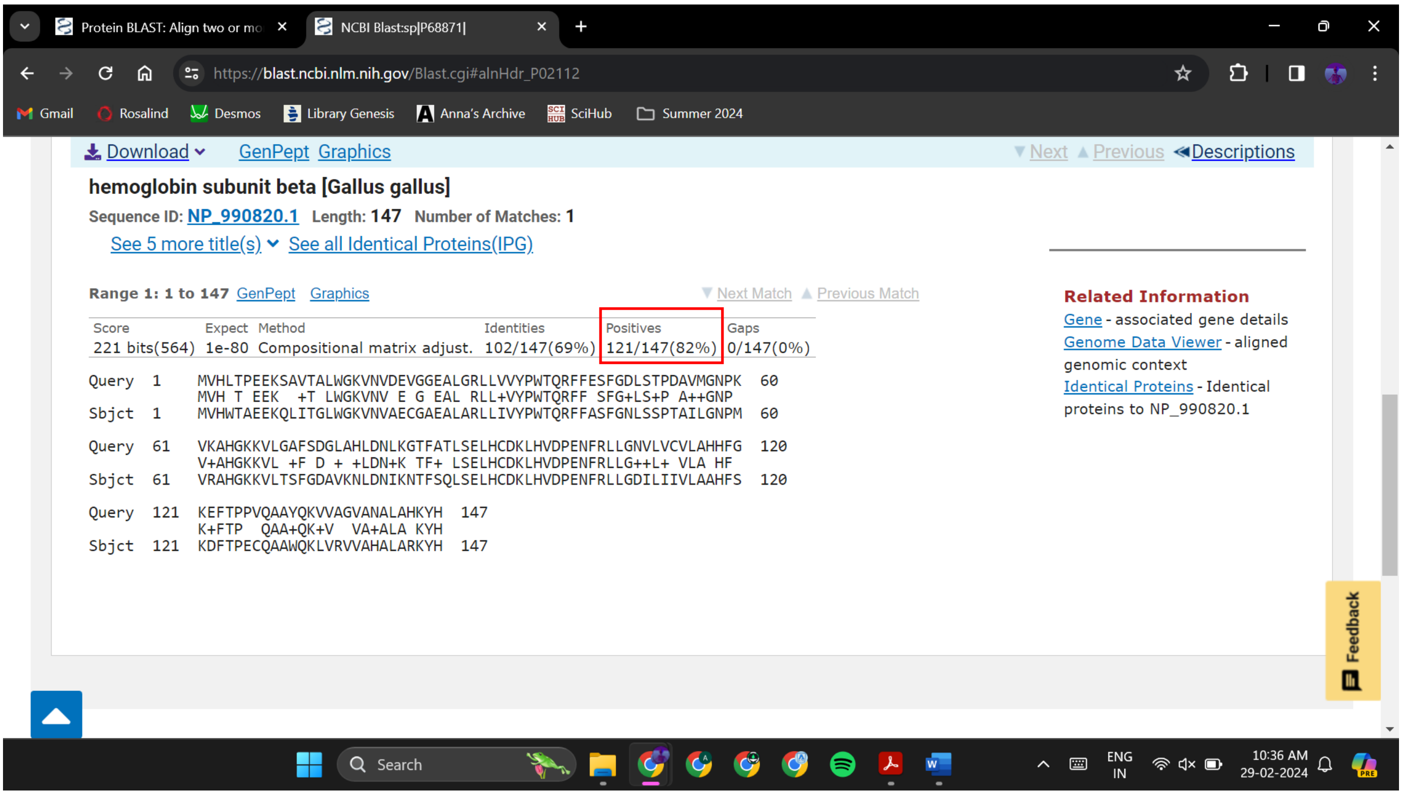
Or, upload file Choose File No file chosen

Program Selection

Algorithm ● blastp (protein-protein BLAST)

Choose a BLAST algorithm

The percentage similarity was found to be 82%.



hemoglobin subunit beta [Gallus gallus]

Sequence ID: NP_990820.1 Length: 147 Number of Matches: 1

See 5 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 147 GenPept 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    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MVH T EEK  +T LWGKVNV E G EAL RLL+VYPWTQRFF SFG+LS+P A++GNP
Sbjct  1    MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPM  60

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

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

Related Information

Gene - associated gene details

Genome Data Viewer - aligned genomic context

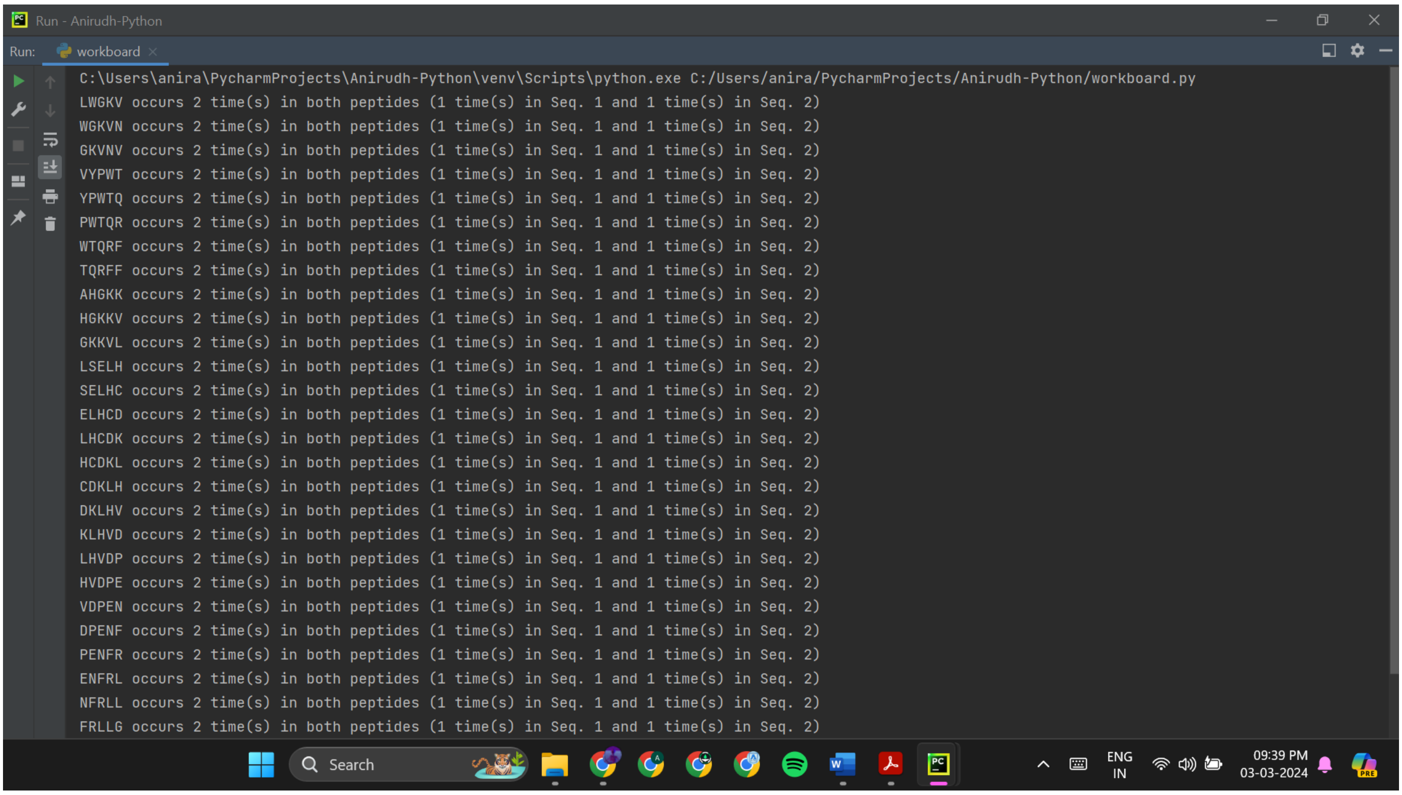
Identical Proteins - Identical proteins to NP_990820.1

The Python code to list all the matching pentapeptides (which occur in both the human and chicken hemoglobin beta sequences) and their frequency of occurrence in given sequences is given below:

```
def find_pentapeptides(seq1, seq2):
    seq1_pentapeptides = {}

    for i in range(len(seq1) - 5 + 1):
        pentapeptide = seq1[i:i + 5]
        if pentapeptide not in seq1_pentapeptides.keys():
            seq1_pentapeptides[pentapeptide] = 1
        else:
            seq1_pentapeptides[pentapeptide] += 1

    matched_pentapeptides = {}
    for i in range(len(seq2) - 5 + 1):
        pentapeptide = seq2[i:i + 5]
        if pentapeptide in seq1_pentapeptides.keys():
            if pentapeptide not in matched_pentapeptides.keys():
                matched_pentapeptides[pentapeptide] = seq1_pentapeptides[pentapeptide]
+ 1
            else:
                matched_pentapeptides[pentapeptide] += 1

    for pentapeptide, frequency in matched_pentapeptides.items():
        print(
            f"{pentapeptide} occurs {frequency} time(s) in both peptides
({seq1_pentapeptides[pentapeptide]} time(s) in Seq. 1 and {frequency -
seq1_pentapeptides[pentapeptide]} time(s) in Seq. 2)")
```

```
human =
"MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTF
ATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH"
chicken =
"MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTF
SQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH"
find_pentapeptides(human, chicken)
```

The output looks like:

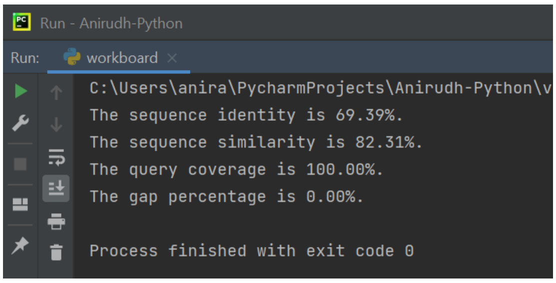


```
C:\Users\anira\PycharmProjects\Anirudh-Python\venv\Scripts\python.exe C:/Users/anira/PycharmProjects/Anirudh-Python/workboard.py
LWGKV occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
WGKVN occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
GKVNV 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)
YPWTQ 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)
WTQRF 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)
AHGKK 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)
GKKVL 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)
SELHC 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)
LHCDK 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)
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)
KLHVD 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)
HVDPE 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)
DPENF 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)
ENFRL 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)
FRLLG occurs 2 time(s) in both peptides (1 time(s) in Seq. 1 and 1 time(s) in Seq. 2)
```

The Python code to compute sequence identity, similarity, query coverage and gap percentage from the alignment of human and chicken hemoglobin sequences is given below:

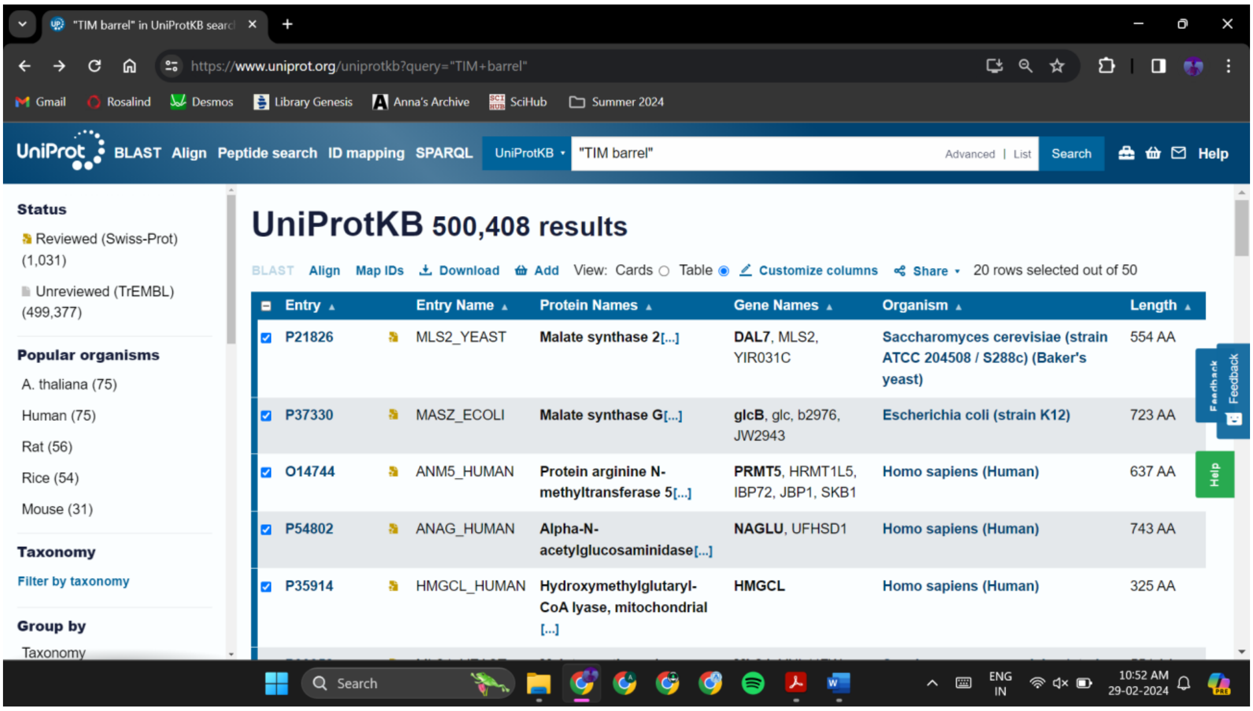
```
def alignment_statistics(seq1, seq2, align):
    identity = 0
    similarity = 0
    gaps = 0

    for char in align:

        if char.isalpha():
            identity += 1
            similarity += 1
        if char == "+":
            similarity += 1
        if char == "-":
            gaps += 1

    alignment_identity = 100 * identity / len(align)
    alignment_similarity = 100 * similarity / len(align)
    query_coverage = 100 * len(align) / len(seq2)
    gap_percentage = 100 * gaps / len(align)

    print(f"The sequence identity is {alignment_identity:.2f}%.")
    print(f"The sequence similarity is {alignment_similarity:.2f}%.")
    print(f"The query coverage is {query_coverage:.2f}%.")
    print(f"The gap percentage is {gap_percentage:.2f}%.")


human =
"MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTF
ATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH"
chicken =
"MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTF
SQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH"
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"

alignment_statistics(human, chicken, alignment)
```

The output looks like:



```
C:\Users\anira\PycharmProjects\Anirudh-Python\v
The sequence identity is 69.39%.
The sequence similarity is 82.31%.
The query coverage is 100.00%.
The gap percentage is 0.00%.

Process finished with exit code 0
```

20 sequences with a TIM barrel domain were selected using UniProt and their sequences were obtained in FASTA format:



Status
Reviewed (Swiss-Prot) (1,031)
Unreviewed (TrEMBL) (499,377)

Popular organisms
A. thaliana (75)
Human (75)
Rat (56)
Rice (54)
Mouse (31)

Taxonomy
Filter by taxonomy

Group by
Taxonomy

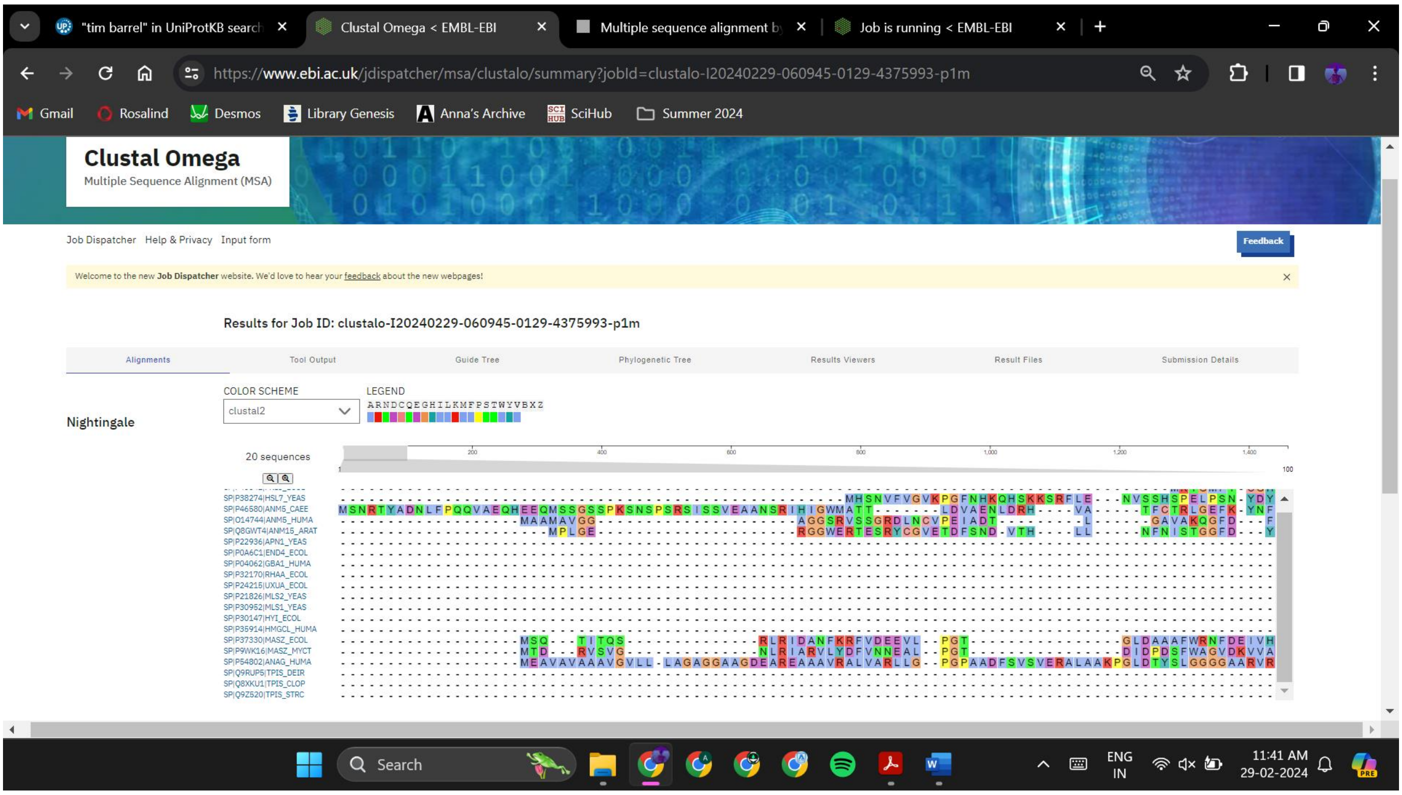
UniProtKB 500,408 results

BLAST Align Map IDs Download Add View: Cards ○ Table ● Customize columns Share 20 rows selected out of 50

☑	Entry		Entry Name	Protein Names	Gene Names	Organism	Length
☑	P21826		MLS2_YEAST	Malate synthase 2[...]	DAL7, MLS2, YIR031C	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	554 AA
☑	P37330		MASZ_ECOLI	Malate synthase G[...]	glcB, glc, b2976, JW2943	Escherichia coli (strain K12)	723 AA
☑	O14744		ANM5_HUMAN	Protein arginine N-methyltransferase 5[...]	PRMT5, HRMT1L5, IBP72, JBP1, SKB1	Homo sapiens (Human)	637 AA
☑	P54802		ANAG_HUMAN	Alpha-N-acetylglucosaminidase[...]	NAGLU, UFHSD1	Homo sapiens (Human)	743 AA
☑	P35914		HMGCL_HUMAN	Hydroxymethylglutaryl-CoA lyase, mitochondrial [...]	HMGCL	Homo sapiens (Human)	325 AA

These were then aligned using Clustal Omega, MAFFT, and MUSCLE:

Clustal Omega

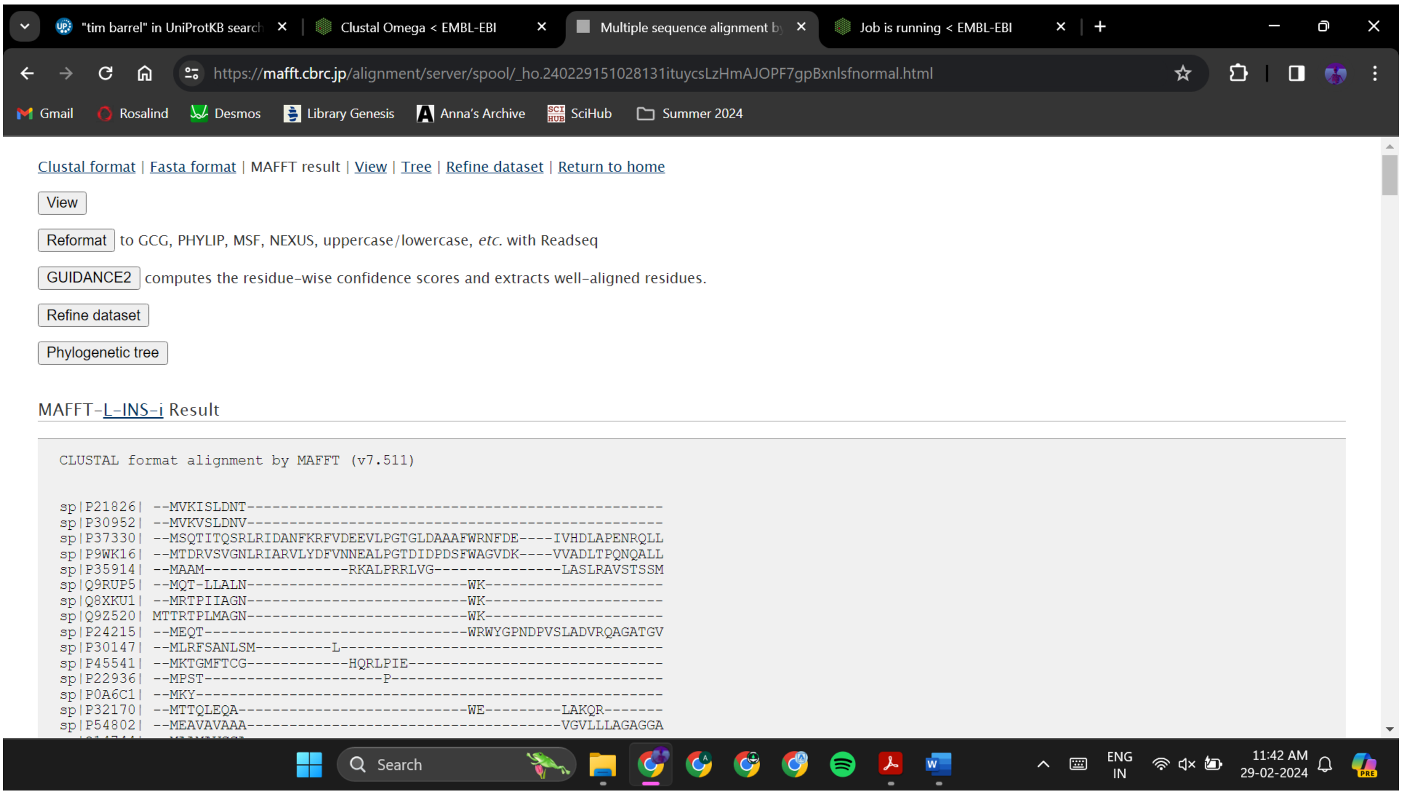


Welcome to the new **Job Dispatcher** website. We'd love to hear your feedback about the new webpages!

Results for Job ID: clustalo-I20240229-060945-0129-4375993-p1m

```
SP|P38274|HSL7_YEAS
SP|P46580|ANM5_CAEE
SP|O14744|ANM5_HUMA
SP|Q8GWT4|ANM15_ARAT
SP|P22936|APN1_YEAS
SP|P0A6C1|END4_ECOL
SP|P04062|GBA1_HUMA
SP|P32170|RHAA_ECOL
SP|P24215|UXUA_ECOL
SP|P21826|MLS2_YEAS
SP|P30952|MLS1_YEAS
SP|P30147|HYI_ECOL
SP|P35914|HMGCL_HUMA
SP|P37330|MASZ_ECOL
SP|P9WK16|MASZ_MYCT
SP|P54802|ANAG_HUMA
SP|Q9RUP5|TPIS_DEIR
SP|Q8XKU1|TPIS_CLOP
SP|Q9Z520|TPIS_STRC
```

MAFFT


"tim barrel" in UniProtKB search
Clustal Omega < EMBL-EBI
Multiple sequence alignment b
Job is running < EMBL-EBI
https://mafft.cbrc.jp/alignment/server/spool/_ho.240229151028131ituycsLzHmAJOPF7gpBxnlsfnormal.html
Gmail
Rosalind
Desmos
Library Genesis
Anna's Archive
SciHub
Summer 2024
View
Reformat
GUIDANCE2
Refine dataset
Phylogenetic tree
Search
ENG
IN
11:42 AM
29-02-2024

Clustal format | Fasta format | MAFFT result | View | Tree | Refine dataset | Return to home

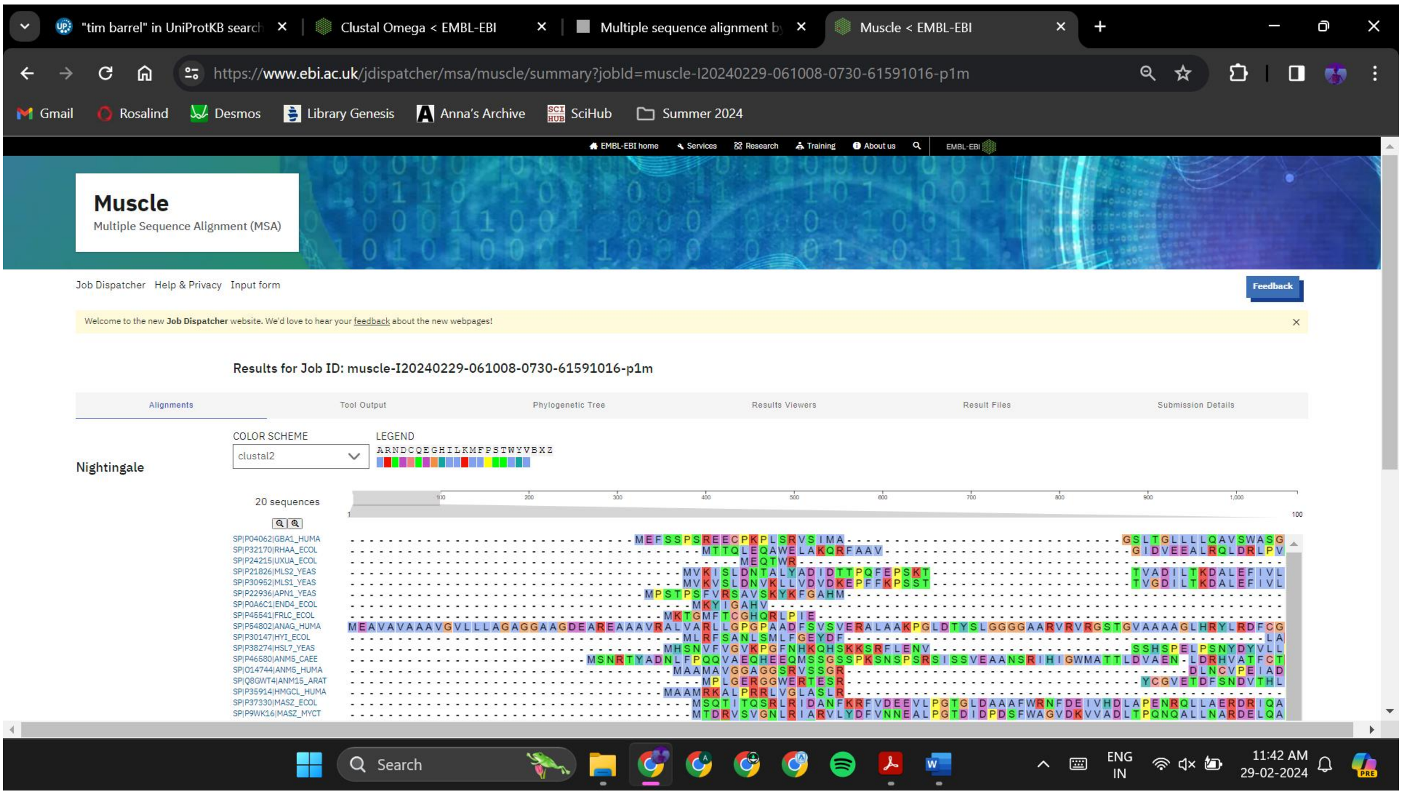
Reformat to GCG, PHYLIP, MSF, NEXUS, uppercase/lowercase, *etc.* with Readseq

GUIDANCE2 computes the residue-wise confidence scores and extracts well-aligned residues.

MAFFT-L-INS-i Result

```
CLUSTAL format alignment by MAFFT (v7.511)


sp|P21826| --MVKISLDNT------------------------------------------------
sp|P30952| --MVKVSLDNV------------------------------------------------
sp|P37330| --MSQTITQSRLRIDANFKRFVDEEVLPGTGLDAAAFWRNFDE----IVHDLAPENRQLL
sp|P9WK16| --MTDRVSVGNLRIARVLYDFVNNEALPGTDIDPDSFWAGVDK----VVADLTPQNQALL
sp|P35914| --MAAM-----------------RKALPRRLVG---------------LASLRAVSTSSM
sp|Q9RUP5| --MQT-LLALN--------------------------WK---------------------
sp|Q8XKU1| --MRTPIIAGN--------------------------WK---------------------
sp|Q9Z520| MTTRTPLMAGN--------------------------WK---------------------
sp|P24215| --MEQT-------------------------------WRWYGPNDPVSLADVRQAGATGV
sp|P30147| --MLRFSANLSM---------L--------------------------------------
sp|P45541| --MKTGMFTCG------------HQRLPIE------------------------------
sp|P22936| --MPST---------------------P--------------------------------
sp|P0A6C1| --MKY-------------------------------------------------------
sp|P32170| --MTTQLEQA---------------------------WE---------LAKQR-------
sp|P54802| --MEAVAVAAA---------------------------------------VGVLLLAGAGGA
```

MUSCLE



"tim barrel" in UniProtKB search | Clustal Omega < EMBL-EBI | Multiple sequence alignment b | Muscle < EMBL-EBI

https://www.ebi.ac.uk/jdispatcher/msa/muscle/summary?jobId=muscle-I20240229-061008-0730-61591016-p1m

Gmail | Rosalind | Desmos | Library Genesis | Anna's Archive | SciHub | Summer 2024

EMBL-EBI home | Services | Research | Training | About us | EMBL-EBI

Muscle
Multiple Sequence Alignment (MSA)

Job Dispatcher | Help & Privacy | Input form | Feedback

Welcome to the new **Job Dispatcher** website. We'd love to hear your feedback about the new webpages!

Results for Job ID: muscle-I20240229-061008-0730-61591016-p1m

Alignments | Tool Output | Phylogenetic Tree | Results Viewers | Result Files | Submission Details

Nightingale

COLOR SCHEME: clustal2

LEGEND: ARNDCQEGHILKMFPSTWYVBXZ

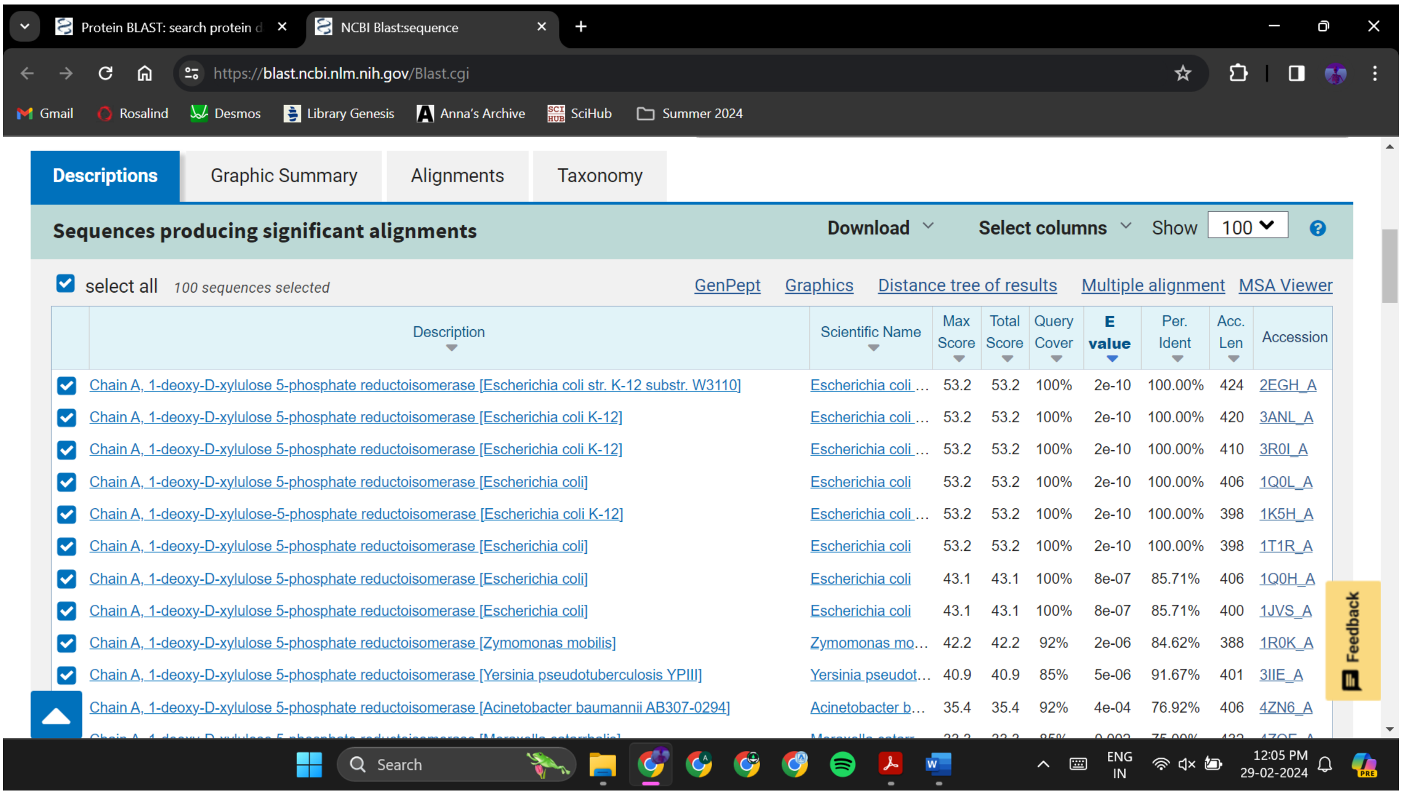
20 sequences

```
SP|P04062|GBA1_HUMA    ------------------------------MEFSSPSREECPKPLSRVSIMA------------------------------GSLTGLLLQAVSWASG
SP|P32170|RHAA_ECOL    -------------------------------------MTTQLEQAWELAKQRFAAV--------------------------GIDVEEALRQLDRLPV
SP|P24215|UXUA_ECOL    -----------------------------------------MEQTWR---------------------------------------------------
SP|P21826|MLS2_YEAS    -----------------------------------MVKISLDNTALYADIDTTPQFEPSKT---------------------TVADILTKDALEFIVL
SP|P30952|MLS1_YEAS    -----------------------------------MVKVSLDNVKLLVDVDKEPFFKPSST---------------------TVGDILTKDALEFIVL
SP|P22936|APN1_YEAS    -------------------------------MPSTPSFVRSAVSKYKFGAHM----------------------------------------------
SP|P0A6C1|END4_ECOL    ------------------------------------MKYIGAHV------------------------------------------------------
SP|P45541|FRLC_ECOL    ---------------------------------MKTGMFTCGHQRLPIE-------------------------------------------------
SP|P54802|ANAG_HUMA    MEAVAVAAAVGVLLLAGAGGAAGDEAREAAAVRALVARLLGPGPAADFSVSVERALAAKPGLDTYSLGGGGAARVRVRGSTGVAAAAGLHRYLRDFCG
SP|P30147|HYI_ECOL     -----------------------------------MLRFSANLSMLFGEYDF--------------------------------------------LA
SP|P38274|HSL7_YEAS    ---------------------------------MHSNVFVGVKPGFNHKQHSKKSRFLENV---------------------SSHSPELPSNYDYVLL
SP|P46580|ANM5_CAEE    -------------------------MSNRTYADNLFPQQVAEQHEEQMSSGSSPKSNSPSRSISSVEAANSRIHIGWMATTLDVAEN-LDRHVATFCT
SP|O14744|ANM5_HUMA    ----------------------------------MAAMAVGGAGGSRVSSGR------------------------------------DLNCVPEIAD
SP|Q8GWT4|ANM15_ARAT   -------------------------------------MPLGERGGWERTESR-------------------------------YCGVETDFSNDVTHL
SP|P35914|HMGCL_HUMA   ---------------------------------MAAMRKALPRRLVGLASLR----------------------------------------------
SP|P37330|MASZ_ECOL    ------------------------------------MSQTITQSRLRIDANFKRFVDEEVLPGTGLDAAAFWRNFDEIVHDLAPENRQLLAERDRIQA
SP|P9WK16|MASZ_MYCT    ------------------------------------MTDRVSVGNLRIARVLYDFVNNEALPGTDIDPDSFWAGVDKVVADLTPQNQALLNARDELQA
```

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

	Clustal Omega	MAFFT	MUSCLE
1	M in P46580, gaps in others	M in Q9Z520, gaps in others	M in P54802, gaps in others
2	S in P46580, gaps in others	T in Q9Z520, gaps in others	E in P54802, gaps in others
3	N in P46580, gaps in others	T in Q9Z520, M in others	A in P54802, gaps in others
4	R in P46580, gaps in others	Different amino acids in different sequences	V in P54802, gaps in others
5	T in P46580, gaps in others	Different amino acids in different sequences	A in P54802, gaps in others

BLASTP was run using the given short sequence with “pdb” as the database:



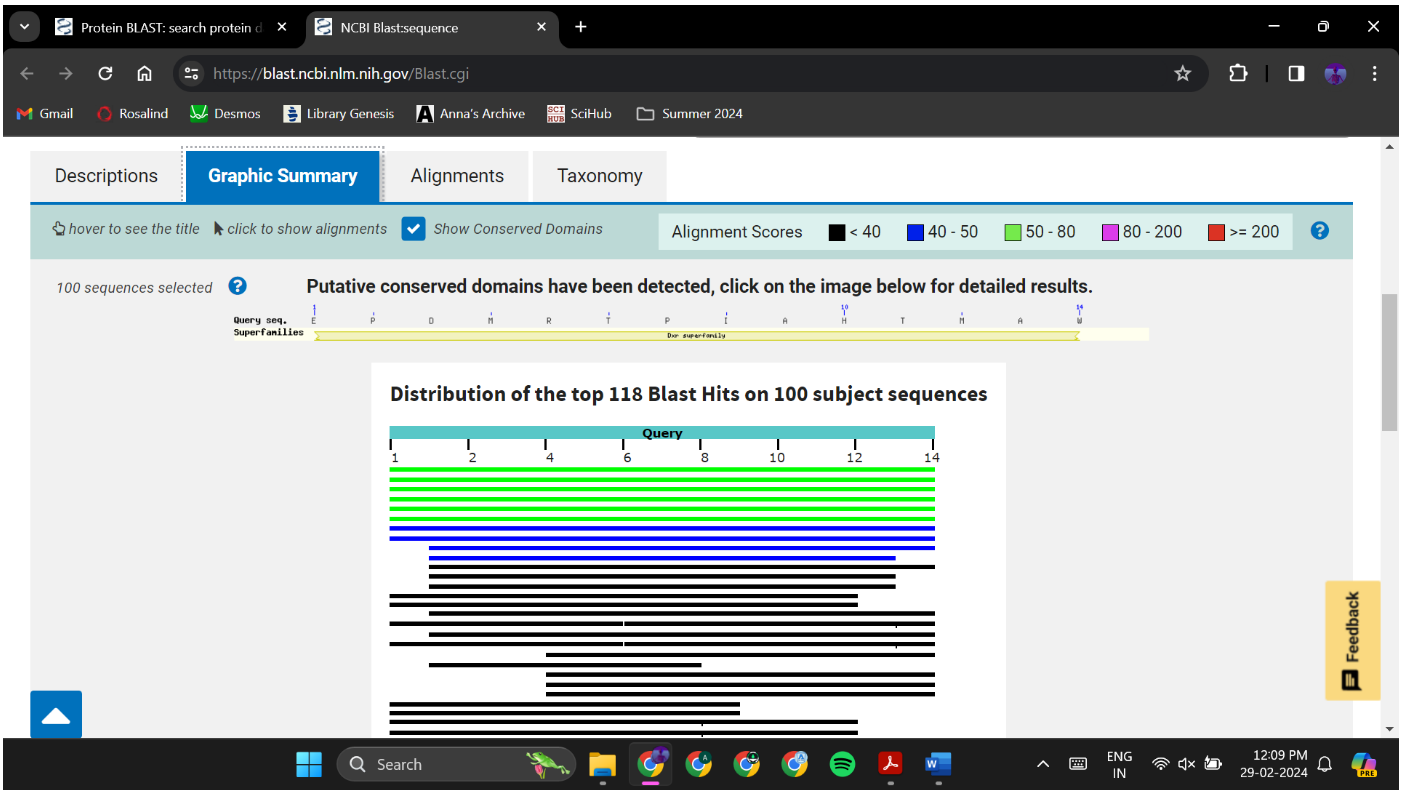
Sequences producing significant alignments

☑ 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 ...	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%	8e-07	85.71%	406	1Q0H_A
☑	Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Escherichia coli]	Escherichia coli	43.1	43.1	100%	8e-07	85.71%	400	1JVS_A
☑	Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Zymomonas mobilis]	Zymomonas mo...	42.2	42.2	92%	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	85%	5e-06	91.67%	401	3IIE_A
	Chain A, 1-deoxy-D-xylulose 5-phosphate reductoisomerase [Acinetobacter baumannii AB307-0294]	Acinetobacter b...	35.4	35.4	92%	4e-04	76.92%	406	4ZN6_A

- Many of the results have 100% query coverage, high percentage identity, and E values very close to 0.
- The query appears to closely match subsequences in Chain A of 1-deoxy-D-xylulose 5-phosphate reductoisomerase in *Escherichia coli* str. K-12.
- From the graphic summary, we can see that the given query sequence is very closely related to the Dxr superfamily:



hover to see the title *click to show alignments* ☑ *Show Conserved Domains*

Alignment Scores ■ < 40 ■ 40 - 50 🟩 50 - 80 ■ 80 - 200 🟥 >= 200

100 sequences selected

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 118 Blast Hits on 100 subject sequences

Just by using a short sequence, we were able to successfully identify a family of related proteins that share the same motif in their sequences.