BT 3040: BIOINFORMATICS

Assignment 5

Atharva Mandar Phatak | BE21B009

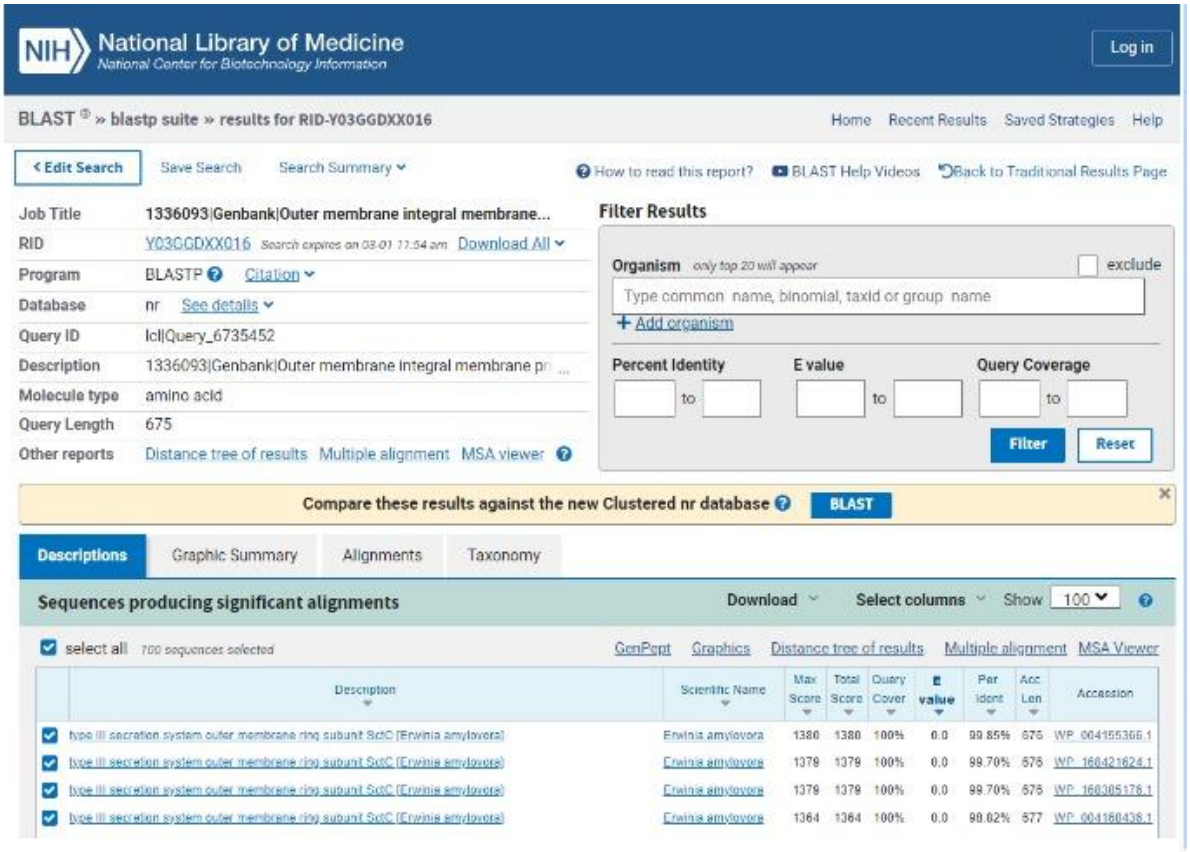
Indian Institute of Technology
Madras

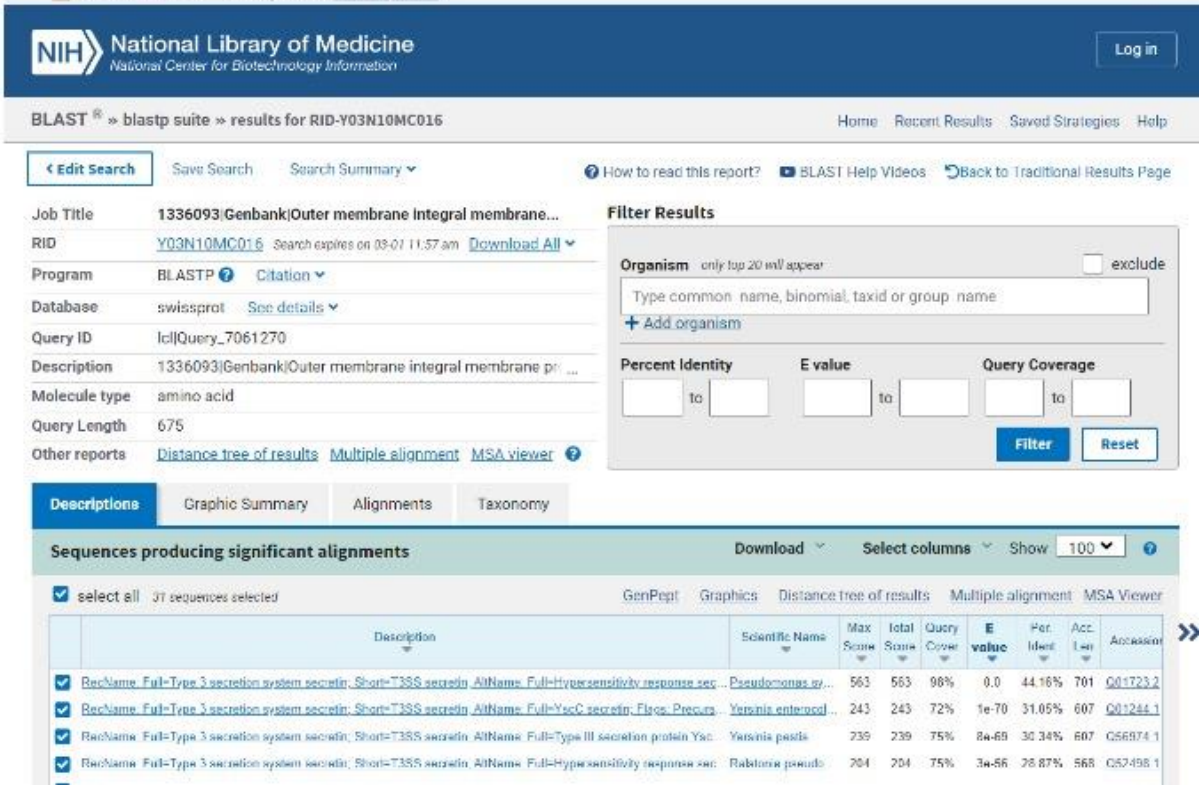
Q1) Analyze the occurrence of similar proteins in "nr" and SWISS-PROT database for the sequence given below:

The following tables provide the analysis of occurrence of similar protiens.

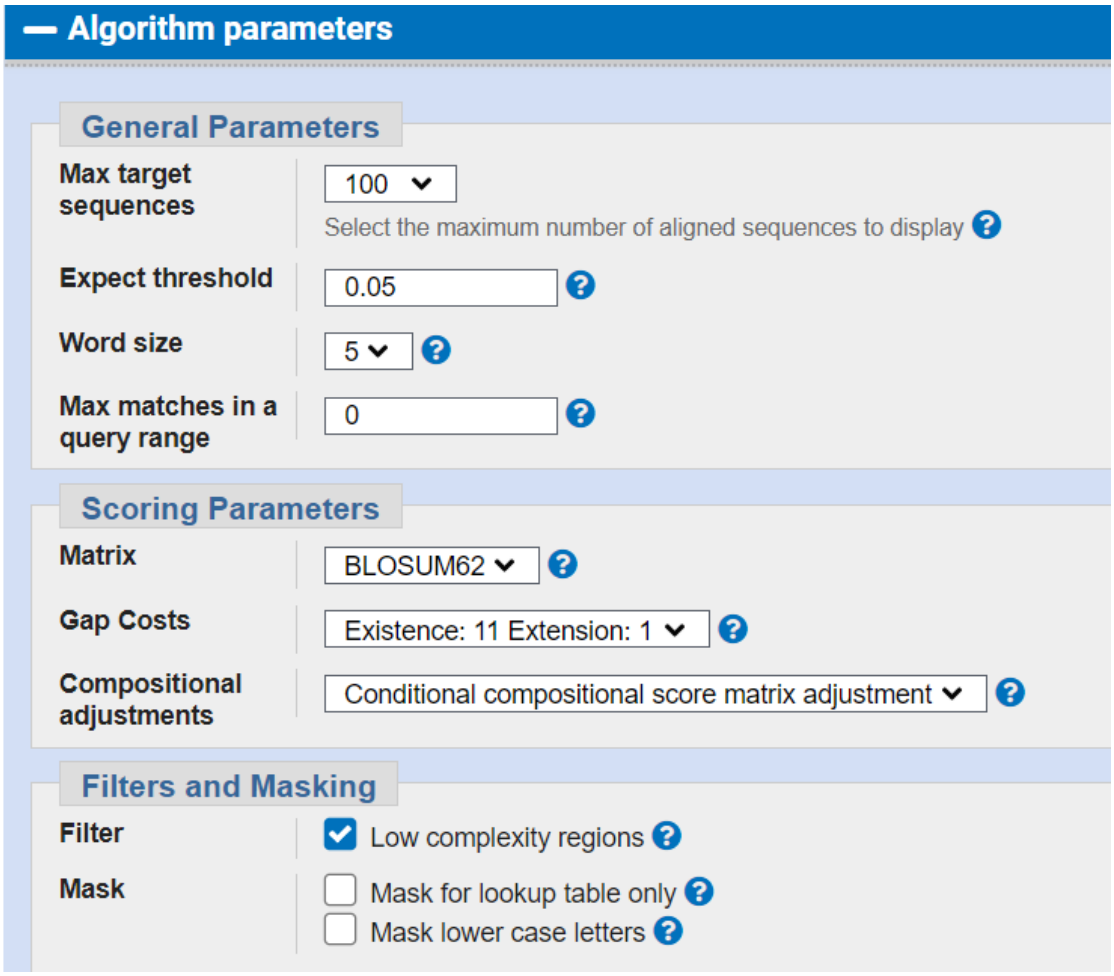
nr		
	Min	Max
Max Score	926	1387
Total Score	1387	1387
Query Cover	90%	100%
e value	0	0
Percentage Identity	66.96	100%
Accession Length	617	676

SwissProt		
	Min	Max
Max Score	42	566
Total Score	42	566
Query Cover	23%	98%
e value	0	0.017
Percentage Identity	21.75	44.16
Accession Length	234	777


National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST ® » blastp suite » results for RID-Y03GGDXX016
Home Recent Results Saved Strategies Help
< Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title 1336093|Genbank|Outer membrane integral membrane...
RID Y03GGDXX016 Search expires on 03-01 11:54 am Download All
Program BLASTP Citation
Database nr See details
Query ID lcl|Query_6735452
Description 1336093|Genbank|Outer membrane integral membrane pr ...
Molecule type amino acid
Query Length 675
Other reports Distance tree of results Multiple alignment MSA viewer
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
+ Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Compare these results against the new Clustered nr database
BLAST
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show 100
select all 700 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 1380 1380 100% 0.0 99.85% 676 WP_004155366.1
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]
Erwinia amylovora 1379 1379 100% 0.0 99.70% 676 WP_160421624.1
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]
Erwinia amylovora 1379 1379 100% 0.0 99.70% 676 WP_160305178.1
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]
Erwinia amylovora 1364 1364 100% 0.0 98.02% 677 WP_004160436.1

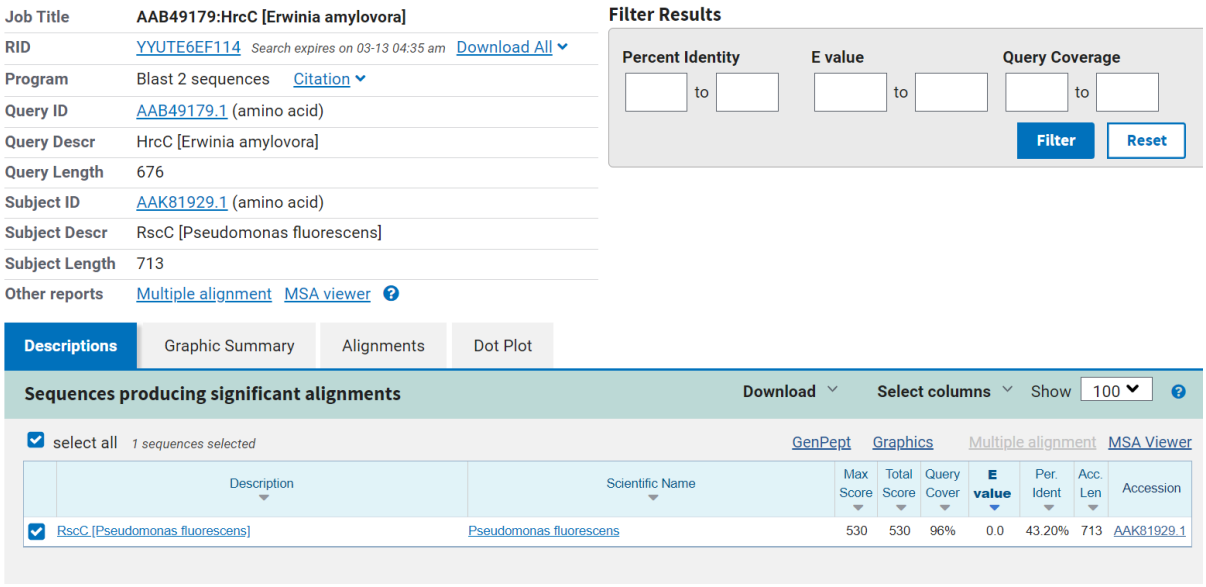

National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST ® » blastp suite » results for RID-Y03N10MC016
Home Recent Results Saved Strategies Help
< Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title 1336093|Genbank|Outer membrane integral membrane...
RID Y03N10MC016 Search expires on 03-01 11:57 am Download All
Program BLASTP Citation
Database swissprot See details
Query ID lcl|Query_7061270
Description 1336093|Genbank|Outer membrane integral membrane pr ...
Molecule type amino acid
Query Length 675
Other reports Distance tree of results Multiple alignment MSA viewer
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
+ Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show 100
select all 37 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... 563 563 98% 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... 243 243 72% 1e-70 31.05% 607 Q01244.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Type III secretion protein Ysc... Yersinia pestis 239 239 75% 8e-69 30.34% 607 Q56974.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Hypersensitivity response sec... Ralstonia pseudo... 204 204 75% 3e-56 28.87% 568 Q57498.1

Q2) List the algorithm parameters used for the search (Q1)



- General parameters displayed:
    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 adjustments
- Filter and Masking
    1. Filter
    2. Mask

Q3) What is the sequence identity of the query sequence (given in Q1) with AAK81929.1



Job Title	AAB49179:HrcC [Erwinia amylovora]
RID	YYUTE6EF114 *Search expires on 03-13 04:35 am* Download All
Program	Blast 2 sequences Citation
Query ID	AAB49179.1 (amino acid)
Query Descr	HrcC [Erwinia amylovora]
Query Length	676
Subject ID	AAK81929.1 (amino acid)
Subject Descr	RscC [Pseudomonas fluorescens]
Subject Length	713
Other reports	Multiple alignment MSA viewer

Filter Results

Percent Identity to | E value to | Query Coverage to | Filter | Reset

Descriptions | Graphic Summary | Alignments | Dot Plot

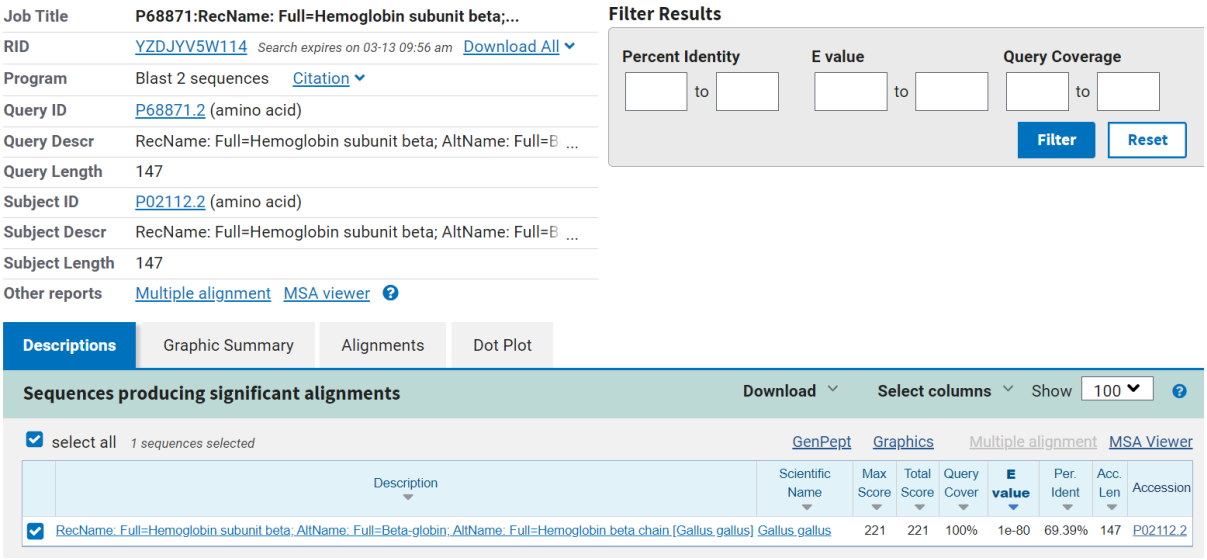
Sequences producing significant alignments — Download | Select columns | Show 100

select all *1 sequences selected* — GenPept | Graphics | Multiple alignment | MSA Viewer

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

- Max Score = 530
- Total Score = 530
- Query Cover = 96%
- E Value = 0
- **Per. Ident = 43.20%**
- Acc. Len = 713

Q4) How far are hemoglobin (beta) sequences in humans and chicken similar?



Job Title	P68871:RecName: Full=Hemoglobin subunit beta;...
RID	YZDJYV5W114 *Search expires on 03-13 09:56 am* Download All
Program	Blast 2 sequences Citation
Query ID	P68871.2 (amino acid)
Query Descr	RecName: Full=Hemoglobin subunit beta; AltName: Full=B ...
Query Length	147
Subject ID	P02112.2 (amino acid)
Subject Descr	RecName: Full=Hemoglobin subunit beta; AltName: Full=B ...
Subject Length	147
Other reports	Multiple alignment MSA viewer

Filter Results

Percent Identity to

E value to

Query Coverage to

Filter Reset

Descriptions | Graphic Summary | Alignments | Dot Plot

Sequences producing significant alignments — Download — Select columns — Show 100

select all *1 sequences selected* GenPept Graphics Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin beta chain [Gallus gallus]	Gallus gallus	221	221	100%	1e-80	69.39%	147	P02112.2

When searched for Haemoglobin sequences in UniProt, we get Human and Chicken protein sequences which are analysed in blast and the Percentage Identity is 69.39%. (Where the query cover is 100%). The percentage Identity is **69.39%**

Q5) Write a program to list all the matching pentapeptides (which occur in both the sequences) and their frequency of occurrence in given sequences.

Code: https://colab.research.google.com/drive/1w2RDIZpFqqW5EFqo6AZ7fBpcZ-ZNdsNA?usp=sharing

Output:

The frequency of WTQRF in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of HVDPE in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of PWTQR in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of FRLLG in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of SELHC in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of TQRFF in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of LSELH in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of GKKVL in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of PENFR in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of LHVDP in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of LWGKV in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of AHGKK in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of YPWTQ in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of WGKVN in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of VYPWT in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of KLHVD in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of NFRLL in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of VDPEN in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of CDKLH in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of HCDKL in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of ELHCD in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of DPENF in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of HGKKV in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of ENFRL in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of GKVNV in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of LHCDK in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1
The frequency of DKLHV in Human Hb Sequence is: 1 and in Chicken Hb Sequence is: 1

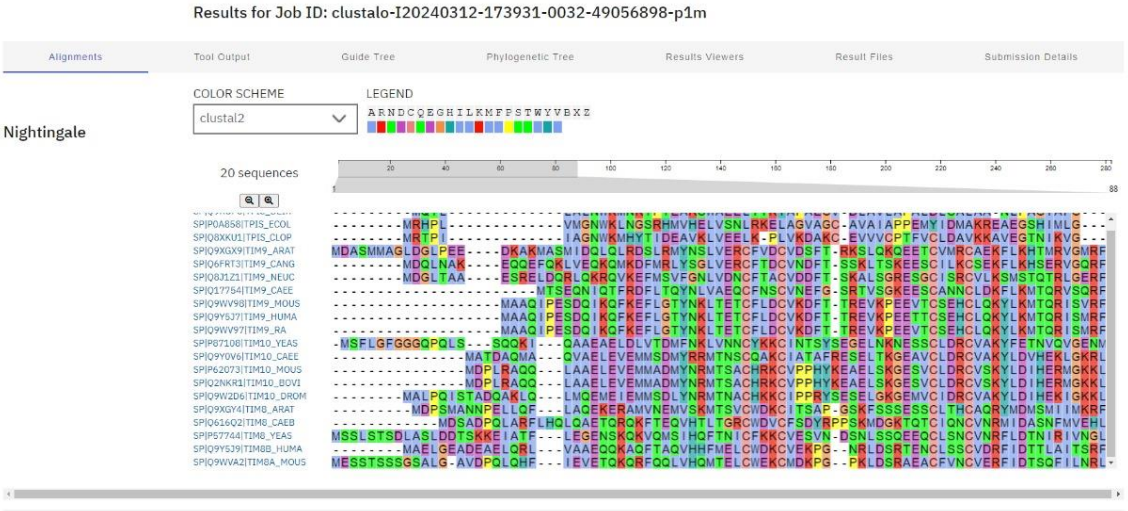
Q6) Write a program to compute sequence identity, similarity, query coverage and gap percentage from the alignment of human and chicken hemoglobin sequences (refer Q4).

```
seq_identity(human,chicken)
69.38775510204081

seq_similarlity(human,chicken,matrix)
82.31292517006803

gap_percentage(human,chicken)
0.0

query_coverage(human,chicken)
query_coverage for human is:  100.0
query_coverage for chicken is:  100.0
```

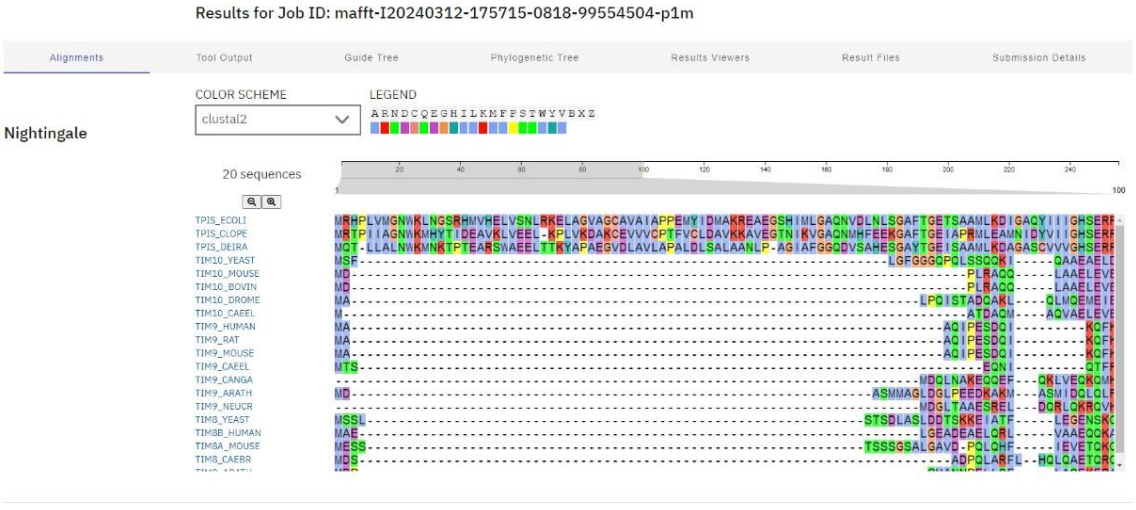
Sequence identity: 69.38775%
Sequence similarity: 82.3129%
Gap percentage: 0%
Query Coverage: 100%

Q7) 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.

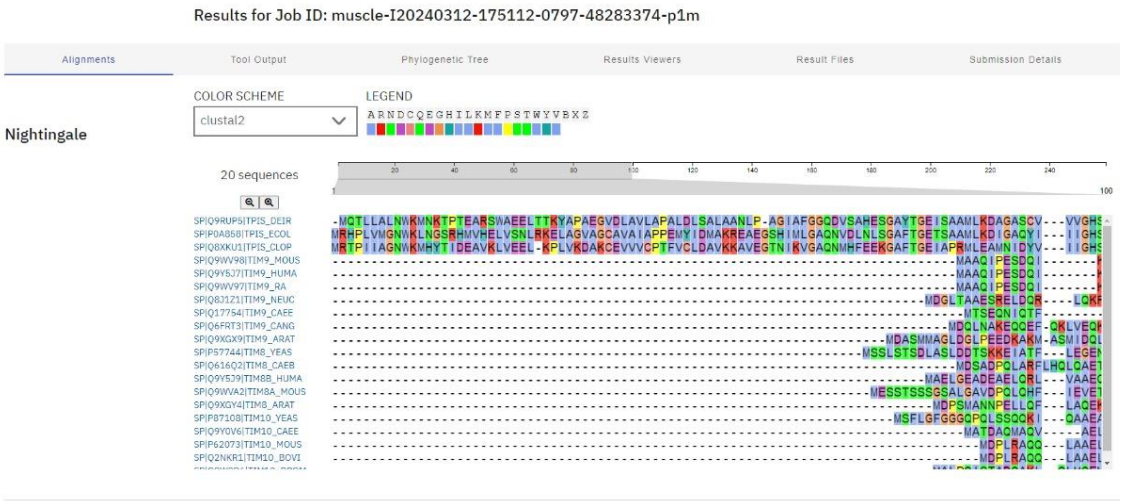
a) Clustal



b) MAFFT



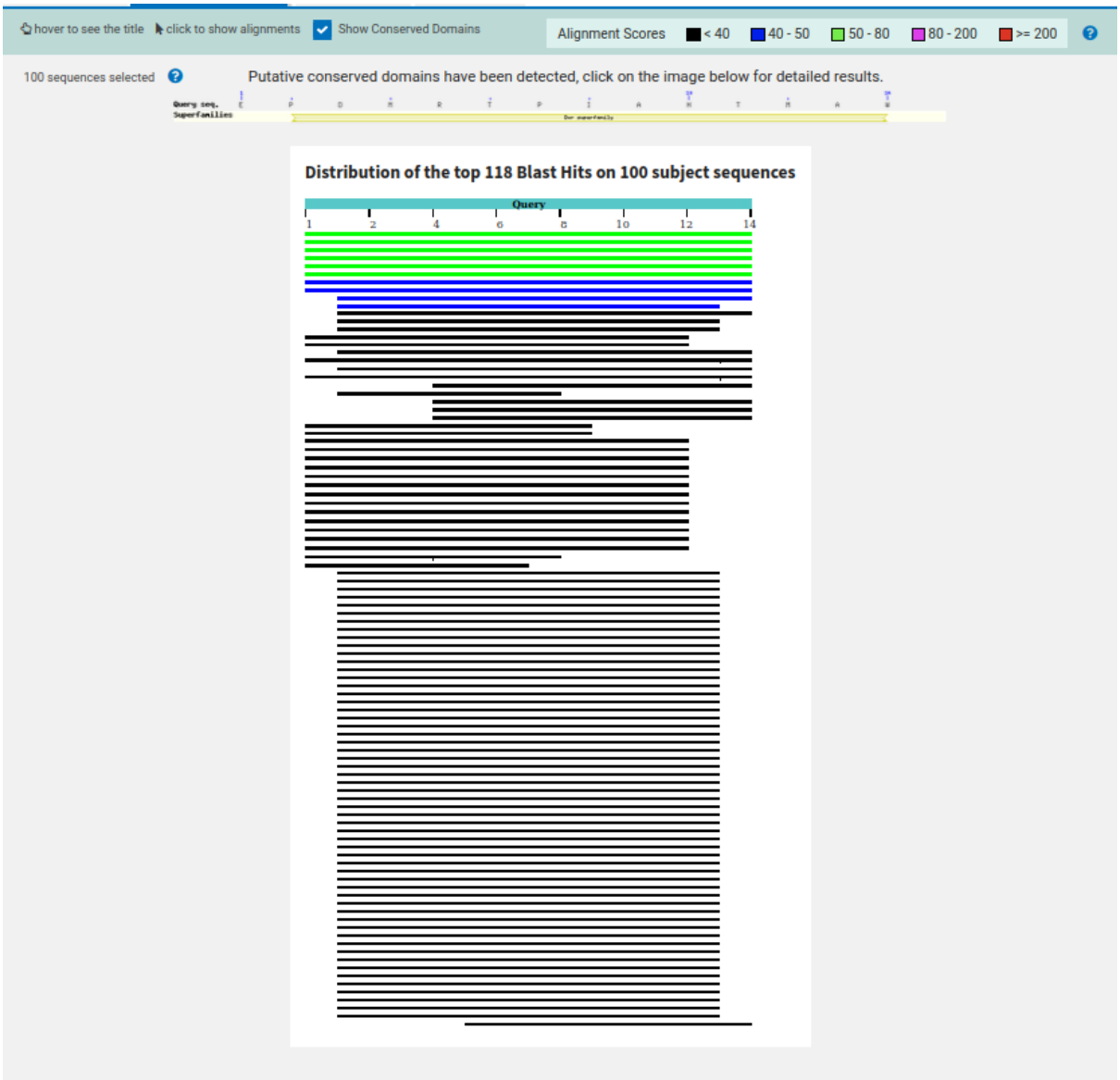
c) MUSCLE



Analysis:

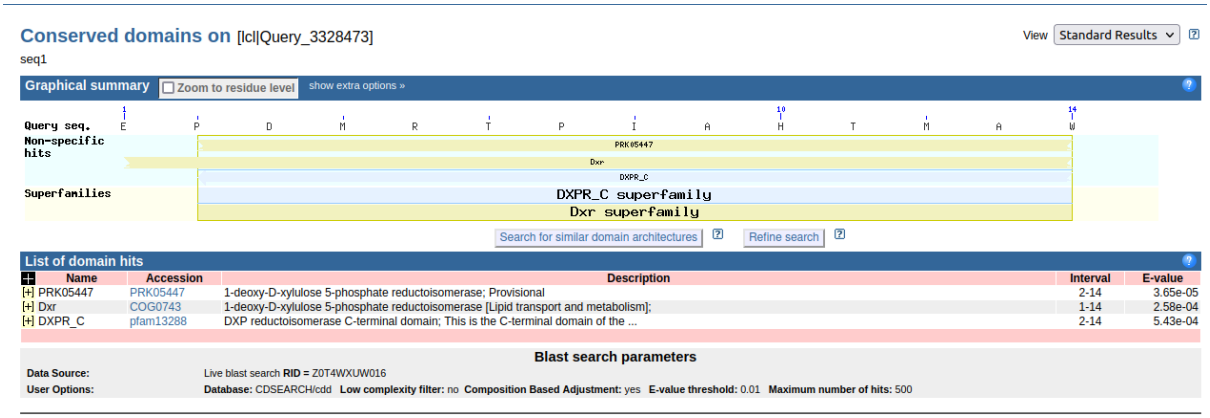
The lengths of the aligned sequences are 120, 117, and 116 for Clustal Omega, MAFFT, and MUSCLE, respectively. It's evident that the alignment of residues at positions 1, 2, 3, and 4 varies across these three methods. As we move further along the residue positions, the alignments become increasingly diverse.

Q8) Blast the below sequence 'EPDMRTPIAHTMAW' against the PDB database. Analyze the results and discuss the significance of the results.



Graphical representation of the BLAST sequence

PDB		
	Min	Max
Max Score	22.30%	53.20%
Total Score	22.30%	53.20%
Query Cover	42%	100%
e value	2.00E-10	21
Percentage Identity	57.14%	100%
Accession Length	201	1290



Thus there are three domain hits for the given sequence.

Domain Hits				
Name	Accession	Description	Interval	E-value
PRK05447	PRK05447	1-deoxy-D-xylulose 5-phosphate reductoisomerase; Provisional	2-14	0.00365%
Dxr	COG0743	1-deoxy-D-xylulose 5-phosphate reductoisomerase [Lipid transport and metabolism];	1-14	0.0258%
DXPR_C	pfam13288	DXP reductoisomerase C-terminal domain; This is the C-terminal domain of the ...	1-14	0.0543%

- Residue 2-14 is a domain named PRK05447 which is conserved across many organisms.
- BLAST results show this sequence is commonly found in - Escherichia coli, Klebsiella pneumoniae.
- 100 sequences, only 10% of the sequences have very minimal E value.