Name:	Mr. Nayan Prabhakar Kasturi
Class:	M. Sc. Bioinformatics (Part I)
Roll Number:	110
Course:	M. Sc. Bioinformatics
Department:	Department of Bioinformatics
Paper:	Mandatory Paper I
Paper Name and Code:	Fundamental of Biology & Bioinformatics (GNKPSBI1501)
Academic Year:	2023-24

SGCP's
Guru Nanak Khalsa College
of Arts, Science & Commerce (Autonomous)

DEPARTMENT OF BIOINFORMATICS

CERTIFICATE

This is to certify that Mr. Nayan Prabhakar Kasturi (Roll No: 110) of M.Sc. Bioinformatics (Part I) has satisfactorily completed the practical for Mandatory Paper 1: Fundamental of Biology & Bioinformatics (GNKPSBI1501) for Semester I course prescribed by the University of Mumbai during the academic year 2023-2024.

TEACHER-IN-CHARGE	**HEAD OF THE DEPARTMENT**	**EXTERNAL EXAMINER**
(Mrs. Aparna Patil Kose)	(Dr. Gursimran Kaur Uppal)	

INDEX

DATE: 01/11/2023

WEBLEM 6
INTRODUCTION TO SEQUENCE ALIGNMENT TOOLS

INTRODUCTION:

Alignment of biological sequences is a fundamental task in bioinformatics. It involves identifying regions of similarity between two or more sequences, which can then be used to infer functional, structural, or evolutionary relationships. Sequence alignment is the problem of comparing biological sequences by searching for a series of nucleotides or amino acids that appear in the same order in the input sequences, possibly introducing gaps into them. When there are two sequences, it is called pairwise sequence alignment; otherwise, it is called multiple sequence alignment (MSA). Global alignment is to find the best match between the entire sequences.

Most MSA methods are based on one of the two pairwise alignment algorithms: the optimal algorithm proposed by Needleman and Wunsch (NW) for global alignment, and the improvement to the NW algorithm proposed by Smith and Waterman (SW) to obtain the local alignment. Various algorithms are employed for sequence alignment, two prominent ones being the Needleman-Wunsch algorithm and the Smith-Waterman algorithm.

The Needleman-Wunsch algorithm performs global alignment, comparing entire sequences, while the Smith-Waterman algorithm is utilized for local alignment, identifying regions of similarity within sequences. These algorithms form the backbone of sequence alignment studies and are accessible through powerful bioinformatics tools available under EMBOSS (European Molecular Biology Open Software Suite). Both algorithms are composed of three phases: initialization, distance matrix computation and trace back. Nevertheless, they differ in their applied techniques at each phase. There are many different techniques used in sequence alignment methods, such as heuristic algorithms, and dynamic programming. Although they ensure the best alignment, dynamic programming methods (such as Needleman-Wunsch and Smith-Waterman) can be computationally demanding for longer sequences. For big datasets, heuristic approaches frequently yield near-optimal alignments, by favoring optimality for of speed and efficiency.

Among the widely used tools and methods, BLAST (Basic Local Alignment Search Tool) and FASTA (Fast Alignment Search Tool) are pivotal in bioinformatics. BLAST uses heuristic methods for comparing sequences quickly and efficiently against large databases, allowing rapid identification of homologous sequences. FASTA combines heuristic methods with probability models to perform quick sequence alignments and similarity searches. These tools are used by researchers in a wide range of fields to identify homologous sequences, infer evolutionary relationships, identify functional and structural motifs, and design primers and probes.

Pairwise Alignment Tools

Pairwise alignment tools are typically used to identify regions of similarity between two sequences of unknown evolutionary relationship. They work by comparing the two sequences and identifying regions of identical or similar characters. Gaps are inserted between the

characters of the two sequences so that the identical or similar characters are aligned in successive columns.

BLAST:

BLAST (Basic Local Alignment Search Tool) is a family of sequence alignment algorithms and programs designed to search for regions of similarity between biological sequences. It is used to search for homologous sequences in a database of known sequences, which can be used to identify genes, infer evolutionary relationships, and design primers and probes. It works by comparing a query sequence to a database of sequences using a heuristic approach. This means that it does not search the entire database for matches, but instead uses a number of shortcuts to identify potential matches. The first step in BLAST is to break the query sequence into short segments, called words. The length of the words depends on the type of sequence being searched (e.g., DNA or protein). BLAST then searches the database for sequences that contain the same words as the query sequence. If a match is found, BLAST extends the alignment in both directions to find the longest possible alignment. BLAST calculates a score for each alignment, which is based on the similarity of the two sequences and the presence of gaps. The higher the score, the more similar the two sequences are. BLAST then reports the alignments with the highest scores.

Types of BLAST:

There are five types (variants) of BLAST that are differentiated based on the type of sequence (DNA or protein) of the query and database sequences.

1. **BLASTN** compares a nucleotide query sequence to a nucleotide sequence database.
2. **BLASTP** compares a protein query sequence to a protein sequence database.
3. **BLASTX** compares a nucleotide query sequence to a protein sequence database by translating the query sequence into its six possible reading frames and aligning them with the protein sequences.
4. **TBLASTN** compares a protein query sequence to a nucleotide sequence database by translating the nucleotide sequences in all six reading frames and aligning them with the protein sequence.
5. **TBLASTX** compares a nucleotide query sequence to a nucleotide sequence database by translating the query sequence in all six reading frames and aligning them with the nucleotide sequences.

FASTA:

FASTA (Fast Alignment Search Tool) is a sequence alignment algorithm and program that is used to search for regions of similarity between biological sequences. It works by first building a hash table of the query sequence. The hash table is a data structure that allows FASTA to quickly find all of the sequences in the database that contain the same words as the query sequence. It then aligns the query sequence to each of the matching sequences in the database to find the longest possible alignment. It calculates a score for each alignment, which is based on the similarity of the two sequences and the presence of gaps. The higher the score, the more similar the two sequences are. It then reports the alignments with the highest scores. It is often used in conjunction with BLAST to identify and analyze homologous sequences. FASTA is

also used to design primers and probes for PCR and other molecular biology techniques.

PSI-BLAST:

PSI-BLAST (Position-Specific Iterative BLAST) is a sequence alignment tool that uses a position-specific scoring matrix (PSSM) to search for distant homologs in protein sequences. It is particularly well-suited for identifying homologs that have diverged significantly from their known relatives. It works by first running a regular BLAST search of the protein sequence database using the query sequence. This produces a list of initial hits. It then constructs a PSSM from the alignments of the initial hits. The PSSM is a statistical model that describes the probability of each amino acid at each position in the alignment. PSI-BLAST then uses the PSSM to search the protein sequence database again. This produces a list of new hits. It then repeats this process, using the PSSM from the previous iteration to search for new hits. PSI-BLAST continues to iterate until the PSSM no longer changes or until a certain number of iterations have been reached. PSI-BLAST then reports the alignments with the highest scores.

PHI-BLAST:

PHI-BLAST (Phylogenetically Inconsistent BLAST) is a sequence alignment tool that uses a probabilistic model to search for distant homologs in protein sequences. It is particularly well-suited for identifying homologs that have diverged significantly from their known relatives. It works by first building a phylogenetic tree of the known homologs of the query sequence. It then uses this tree to generate a position-specific scoring matrix (PSSM) for each node in the tree. The PSSM is a statistical model that describes the probability of each amino acid at each position in the alignment. It then searches the database of protein sequences for sequences that match the PSSMs at the nodes of the phylogenetic tree. It does this by calculating a score for each alignment based on the similarity of the sequences and the PSSM. The higher the score, the more similar the sequences are and the more likely they are to be homologous. It then reports the alignments with the highest scores. It also reports the probability that each alignment is a true homolog. This probability is based on the score of the alignment, the PSSM of the node in the phylogenetic tree, and the phylogenetic relationships between the sequences in the alignment. PHI-BLAST is a powerful tool for identifying distant homologs. It is used by researchers in a wide range of fields, including genetics, genomics, proteomics, and molecular biology.

EMBOSS Needle:

EMBOSS Needle is a pairwise sequence alignment tool that uses the Needleman- Wunsch algorithm to produce global alignments. A global alignment is an alignment that aligns the entire length of both sequences. It works by comparing the two sequences and identifying regions of identical or similar characters. Gaps are inserted between the characters of the two sequences so that the identical or similar characters are aligned in successive columns. It calculates a score for each alignment, which is based on the similarity of the two sequences and the presence of gaps. The higher the score, the more similar the two sequences are. It then reports the alignment with the highest score. EMBOSS Needle is a powerful tool for aligning biological sequences and it is particularly well-suited for aligning sequences of known evolutionary relationship or sequences with low levels of divergence.

EMBOSS Water:

EMBOSS Water is a pairwise alignment tool that uses the Smith-Waterman algorithm to produce local alignments. This means that only the most similar regions of the two sequences are aligned. It is a good choice for aligning sequences of unknown evolutionary relationship or sequences with high levels of divergence. It works by comparing the two sequences and identifying regions of identical or similar characters. Gaps are inserted between the characters of the two sequences so that the identical or similar characters are aligned in successive columns. It then calculates a score for each alignment, which is based on the similarity of the two sequences and the presence of gaps. The higher the score, the more similar the two sequences are. It then reports the alignment with the highest score. It is a powerful tool for aligning biological sequences. It is often used in conjunction with other alignment tools, such as BLAST and FASTA, to identify and analyze homologous sequences. EMBOSS Water is also used to design primers and probes for PCR and other molecular biology techniques.

REFERENCES:

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2
2. Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4
3. Bhagwat, M., & Aravind, L. (2007). PSI-BLAST Tutorial. In Methods in molecular biology (pp. 177–186). https://doi.org/10.1007/978-1-59745-514-5_10
4. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., López, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, 7(1). https://doi.org/10.1038/msb.2011.75

DATE: 01/11/2023

WEBLEM 6(A)
BASIC LOCAL ALIGNMENT SEARCH TOOL (BLAST)
(URL: https://blast.ncbi.nlm.nih.gov)

AIM:

To study and explore similar sequences of the protein albumin (UniProt ID: P02768) by using Basic Local Alignment Search Tool (BLAST).

INTRODUCTION:

BLAST (Basic Local Alignment Search Tool) is an algorithm and program for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA and/or RNA sequences. A BLAST search enables a researcher to compare a subject protein or nucleotide sequence (called a query) with a library or database of sequences, and identify database sequences that resemble the query sequence above a certain threshold. BLAST (Basic Local Alignment Search Tool) has become the defacto standard in search and alignment tools [Altschul et al., 1990]. The BLAST algorithm works by finding a short, or local, region of high similarity between two sequences, and then extending this match out from this starting point to both the left and the right. A score is assigned to the match. The score will increase as more residues are found to match and will decrease if there are gaps in the alignment. Alignments with a score that exceeds a certain threshold are reported in the output.

BLAST searches for high scoring sequence alignments between the query sequence and the existing sequences in the database using a heuristic approach that approximates the Smith-Waterman algorithm.

BLAST tool can be used to identify unknown sequences by comparing them with known sequences in a database which helps in predicting the functions of proteins or genes which can be used in phylogenetic analysis as well as in identifying functionally conserved domains within proteins which is important for predicting the functions of proteins.

Albumin:

Albumin is a family of globular proteins, with the most common members being the serum albumins. All proteins within the albumin family are water-soluble, moderately soluble in concentrated salt solutions, and susceptible to heat denaturation. Albumins are commonly present in blood plasma and distinguish themselves from other blood proteins by their lack of glycosylation. Compounds containing albumins are termed albuminoids. Several blood transport proteins share an evolutionary relationship within the albumin family, including serum albumin, alpha-fetoprotein, vitamin D-binding protein, and afamin. This family is exclusively found in vertebrates. In a broader sense, the term "albumins" may refer to other proteins that coagulate under specific conditions.

METHODOLOGY:

1. Open the Homepage of the UniProt database and search for the query of Albumin protein.
2. Select one entry from the results for *Homo sapiens* (UniProt ID: P02768) and download its FASTA sequence in canonical format.
3. Open the homepage of BLAST and select Protein BLAST, i.e., BLASTP.
4. Paste the FASTA sequence in 'Enter Query Sequence' box.
5. Set the desired parameters.
6. Run the BLAST.

OBSERVATIONS:

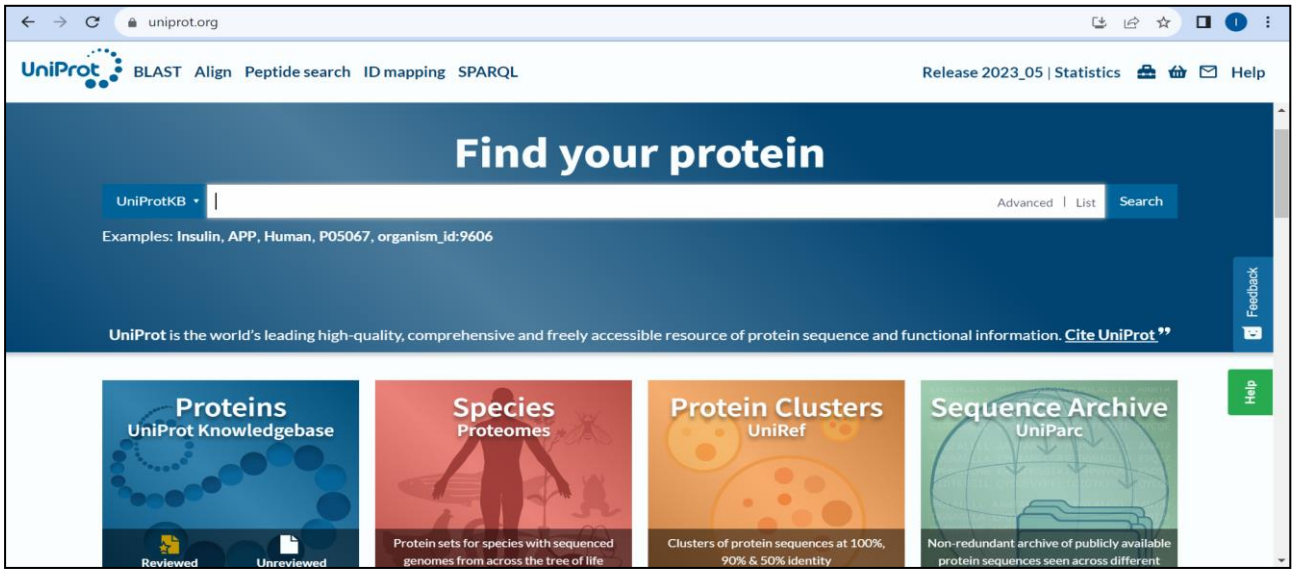


Figure 1: Homepage of the UniProt Database

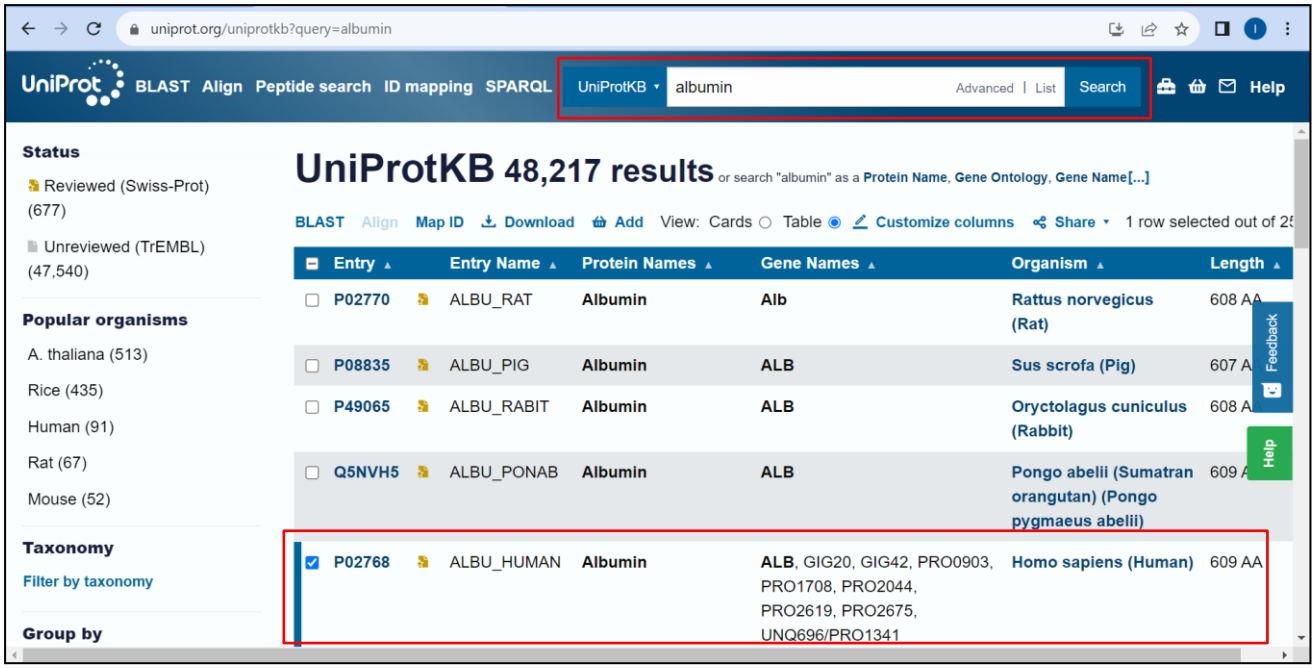


Figure 2: Searching for the query albumin and selecting (UniProt ID: P02768)

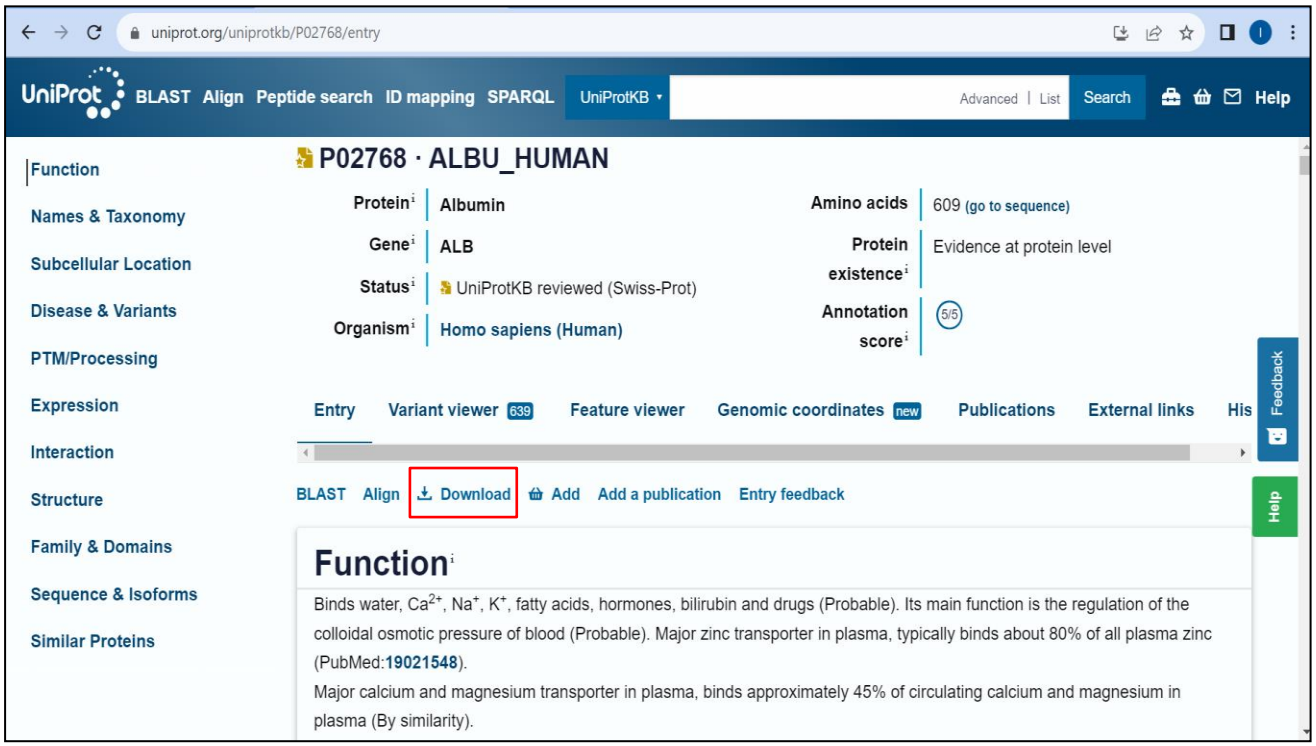


Figure 3: Download option for retrieving FASTA sequence

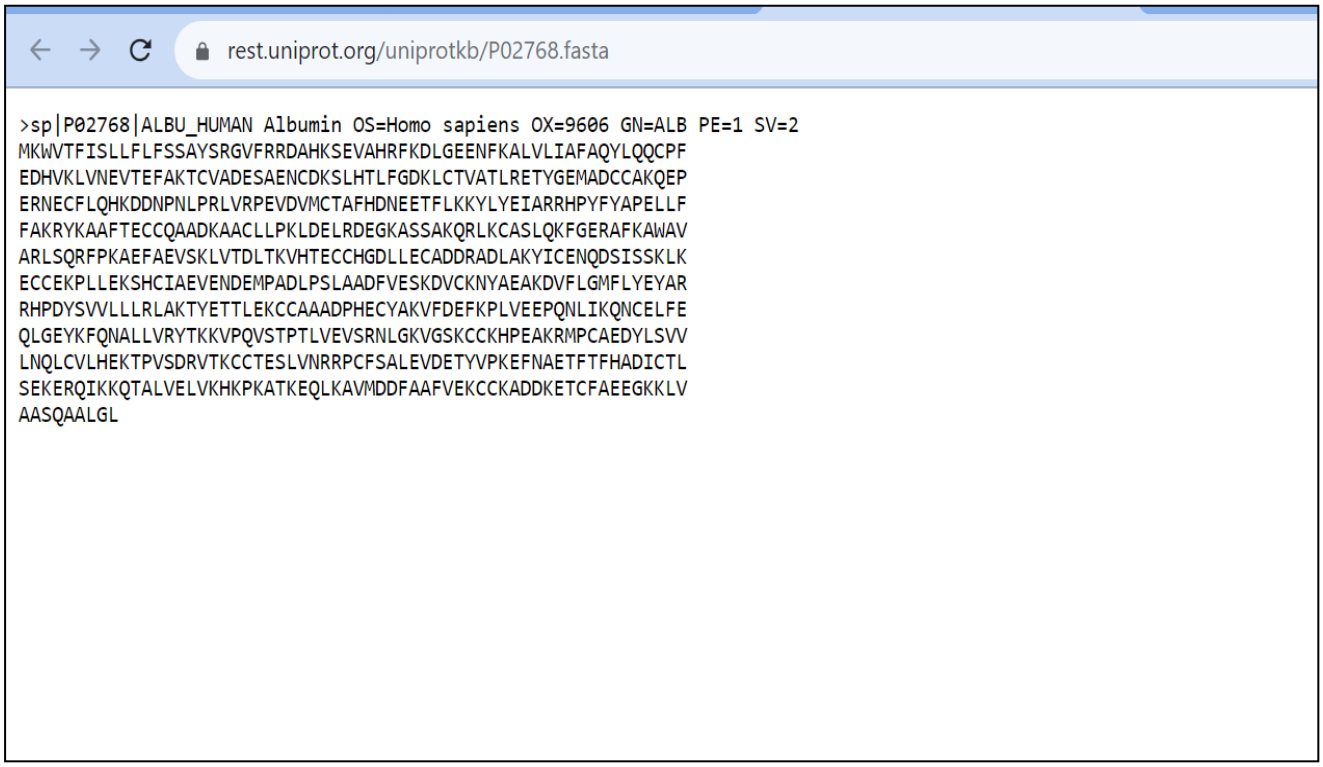


Figure 4: FASTA sequence in canonical format

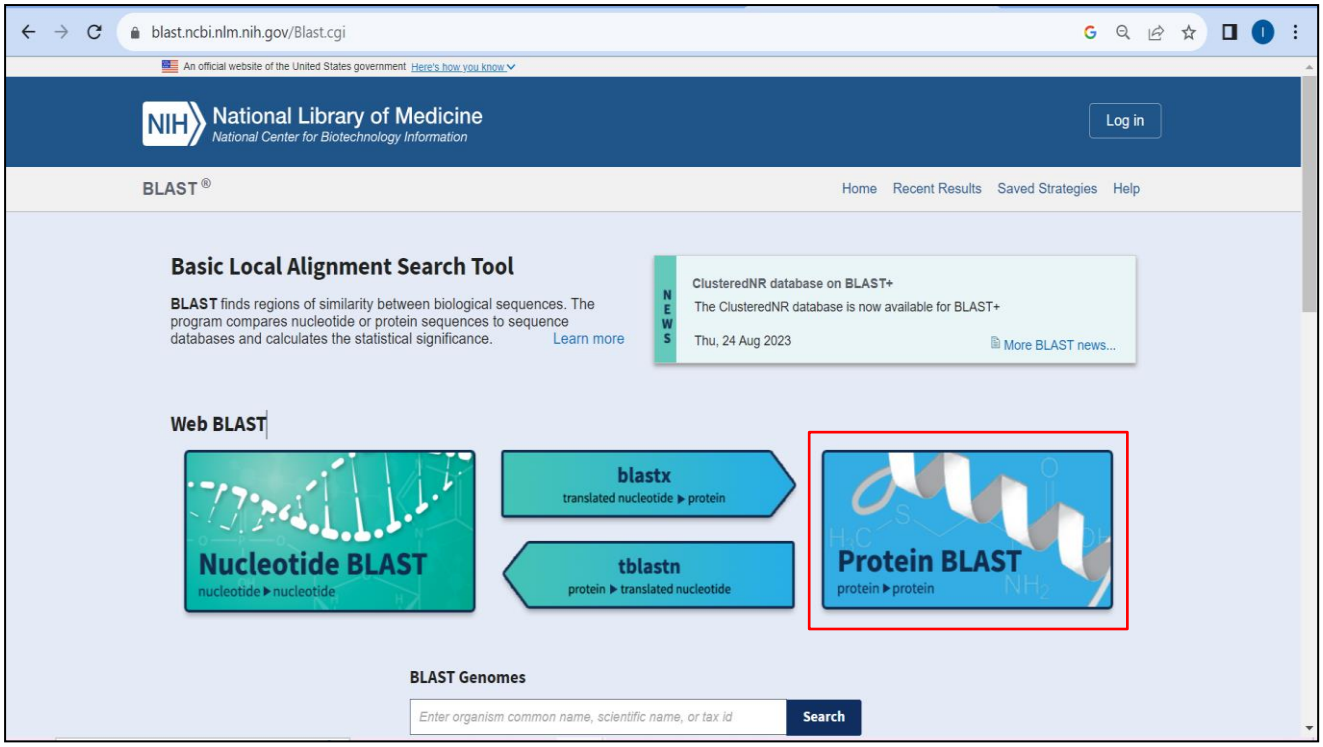


Figure 5: Homepage of Basic Local Alignment Search Tool (BLAST)

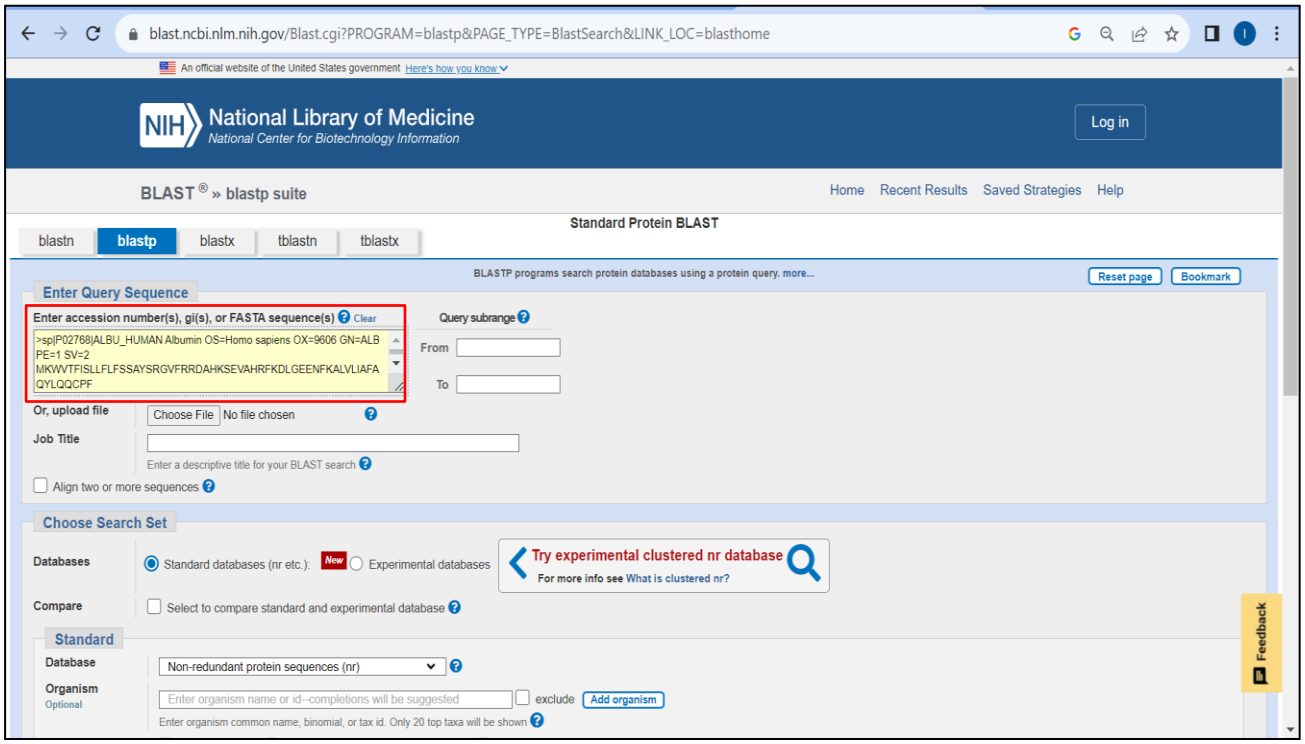


Figure 6: FASTA sequence pasted in 'Enter Query Sequence' box

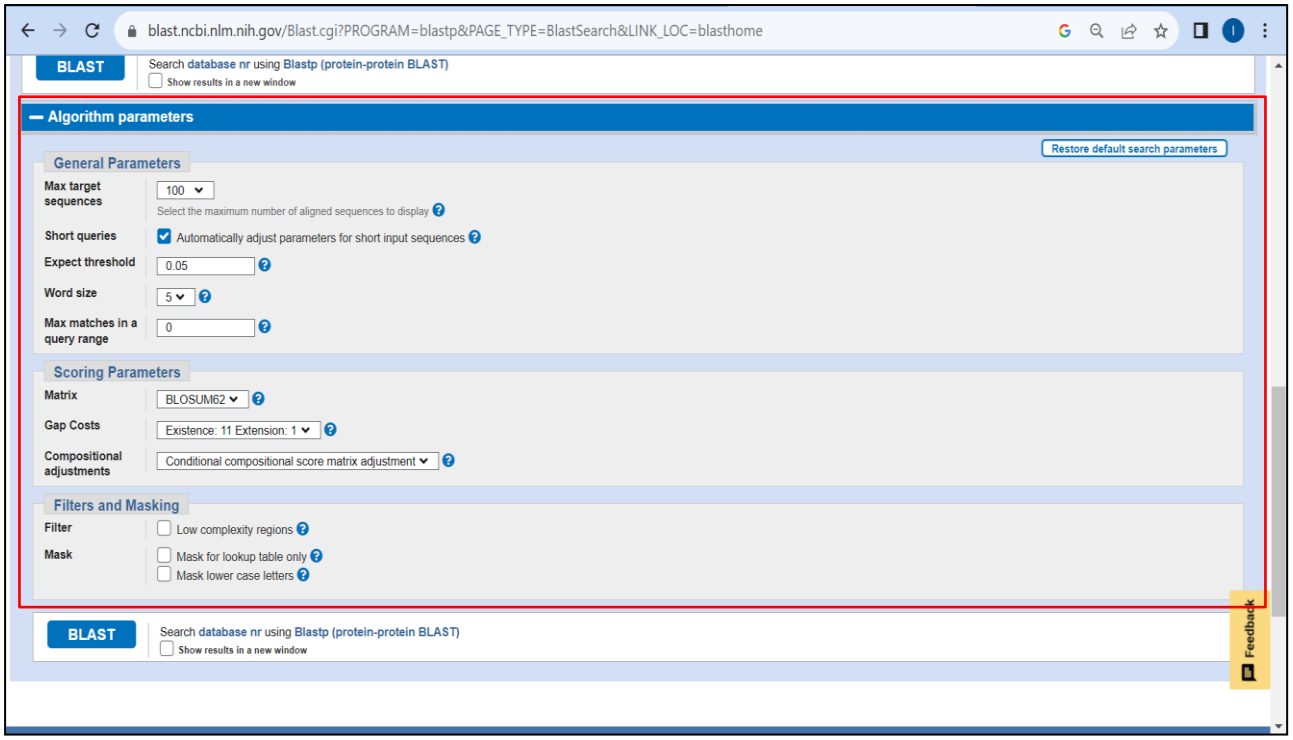


Figure 7: Setting the Algorithm parameters

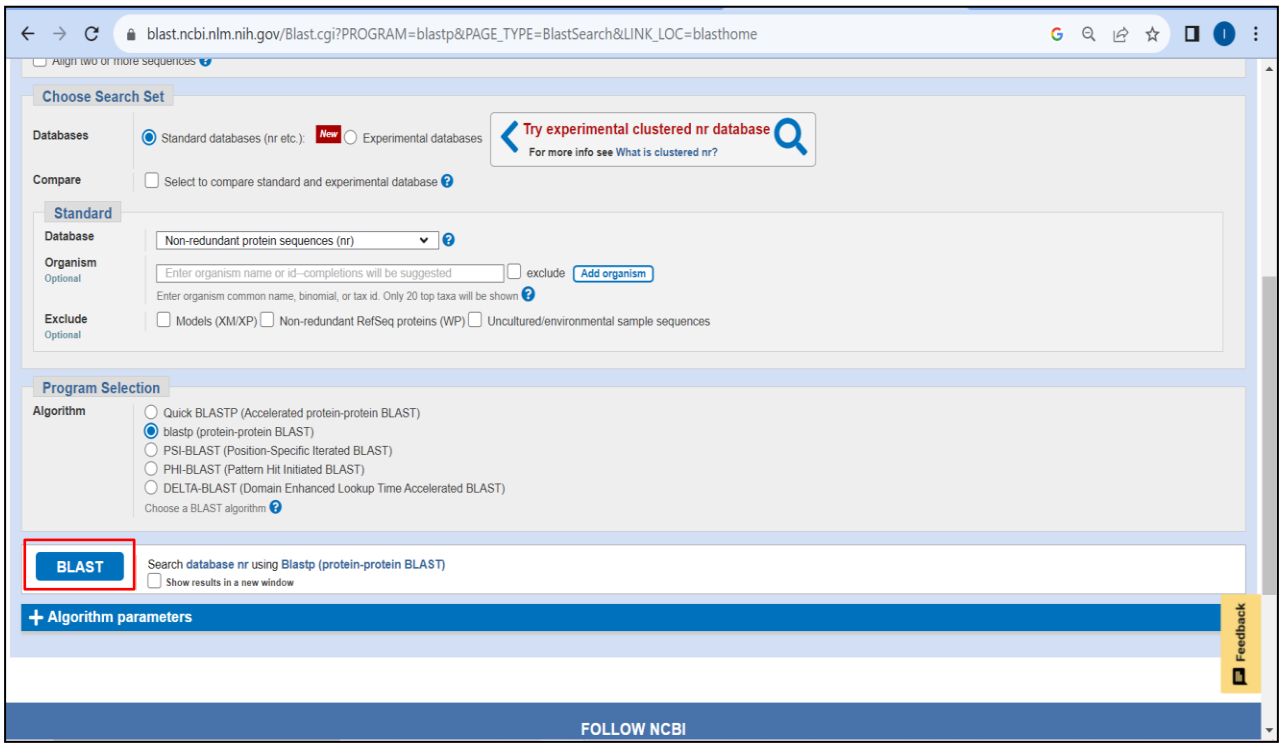


Figure 8: Running BLAST

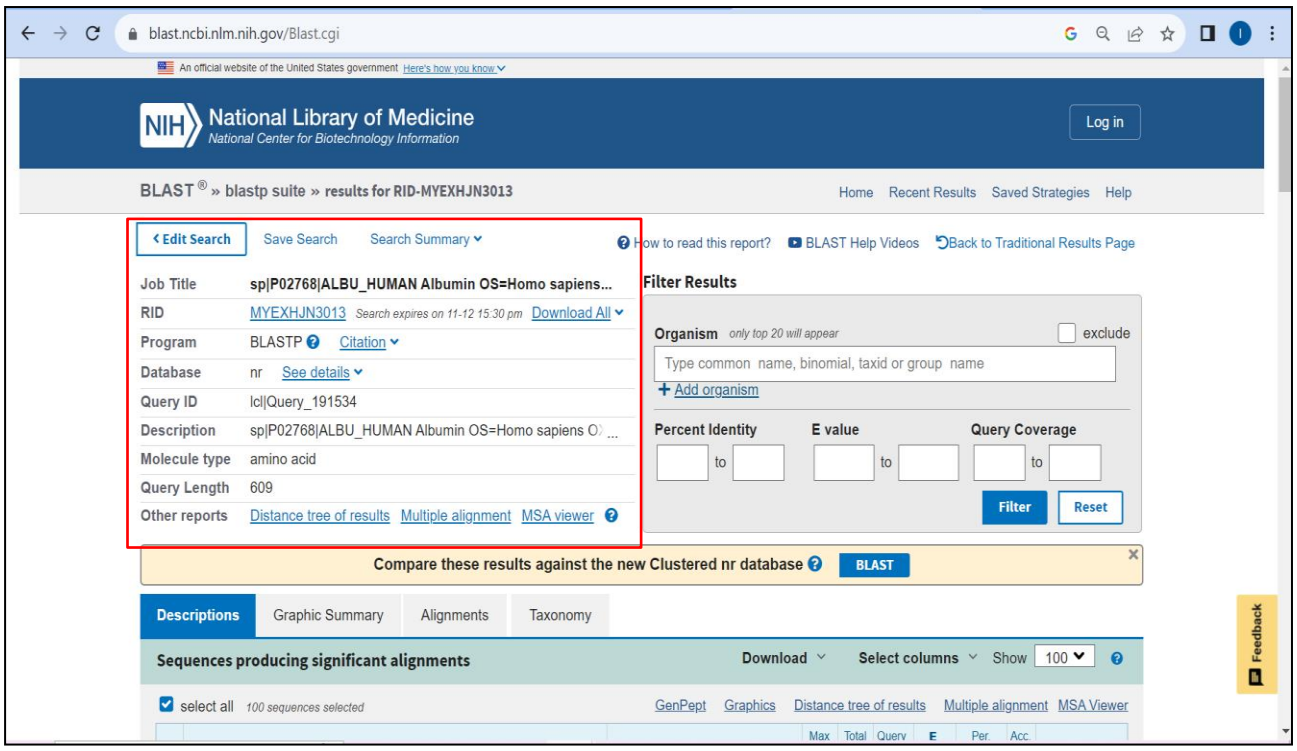
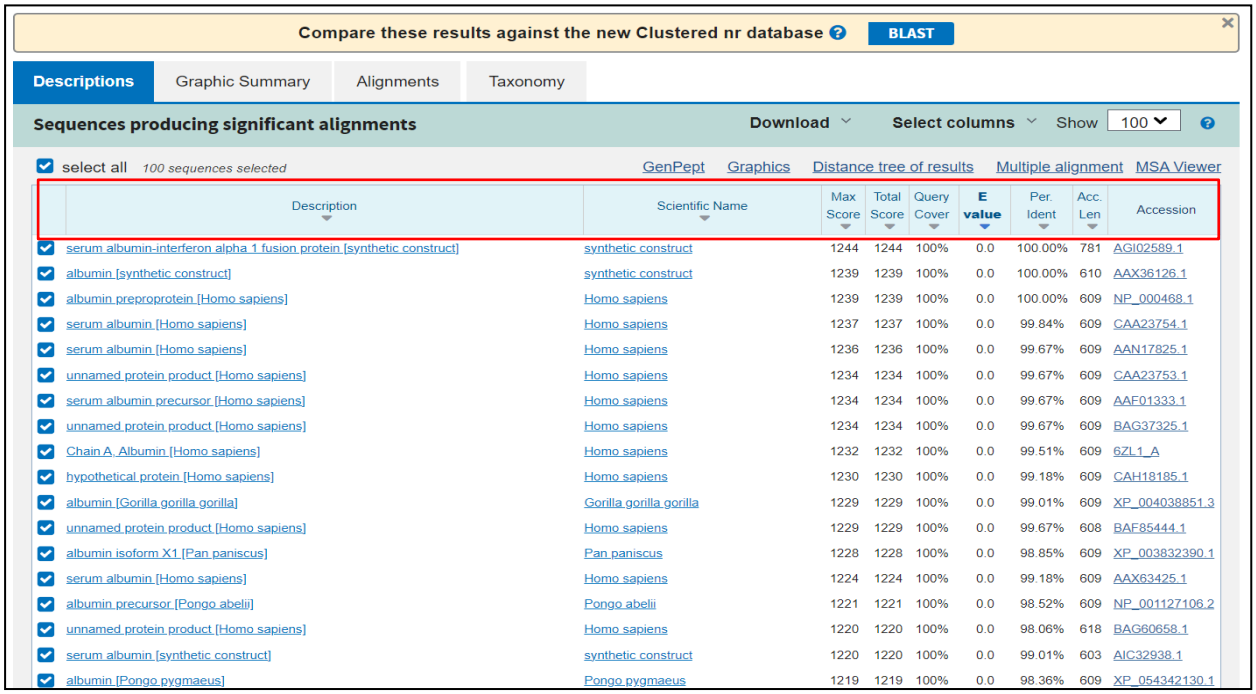


Figure 9: Results for the query, Header Section (UniProt ID: P02768)



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
serum albumin-interferon alpha 1 fusion protein [synthetic construct]	synthetic construct	1244	1244	100%	0.0	100.00%	781	AGI02589.1
albumin [synthetic construct]	synthetic construct	1239	1239	100%	0.0	100.00%	610	AAX36126.1
albumin preproprotein [Homo sapiens]	Homo sapiens	1239	1239	100%	0.0	100.00%	609	NP_000468.1
serum albumin [Homo sapiens]	Homo sapiens	1237	1237	100%	0.0	99.84%	609	CAA23754.1
serum albumin [Homo sapiens]	Homo sapiens	1236	1236	100%	0.0	99.67%	609	AAN17825.1
unnamed protein product [Homo sapiens]	Homo sapiens	1234	1234	100%	0.0	99.67%	609	CAA23753.1
serum albumin precursor [Homo sapiens]	Homo sapiens	1234	1234	100%	0.0	99.67%	609	AAF01333.1
unnamed protein product [Homo sapiens]	Homo sapiens	1234	1234	100%	0.0	99.67%	609	BAG37325.1
Chain A, Albumin [Homo sapiens]	Homo sapiens	1232	1232	100%	0.0	99.51%	609	6ZL1_A
hypothetical protein [Homo sapiens]	Homo sapiens	1230	1230	100%	0.0	99.18%	609	CAH18185.1
albumin [Gorilla gorilla gorilla]	Gorilla gorilla gorilla	1229	1229	100%	0.0	99.01%	609	XP_004038851.3
unnamed protein product [Homo sapiens]	Homo sapiens	1229	1229	100%	0.0	99.67%	608	BAF85444.1
albumin isoform X1 [Pan paniscus]	Pan paniscus	1228	1228	100%	0.0	98.85%	609	XP_003832390.1
serum albumin [Homo sapiens]	Homo sapiens	1224	1224	100%	0.0	99.18%	609	AAX63425.1
albumin precursor [Pongo abelii]	Pongo abelii	1221	1221	100%	0.0	98.52%	609	NP_001127106.2
unnamed protein product [Homo sapiens]	Homo sapiens	1220	1220	100%	0.0	98.06%	618	BAG60658.1
serum albumin [synthetic construct]	synthetic construct	1220	1220	100%	0.0	99.01%	603	AIC32938.1
albumin [Pongo pygmaeus]	Pongo pygmaeus	1219	1219	100%	0.0	98.36%	609	XP_054342130.1

Figure 10: Result for Description Section

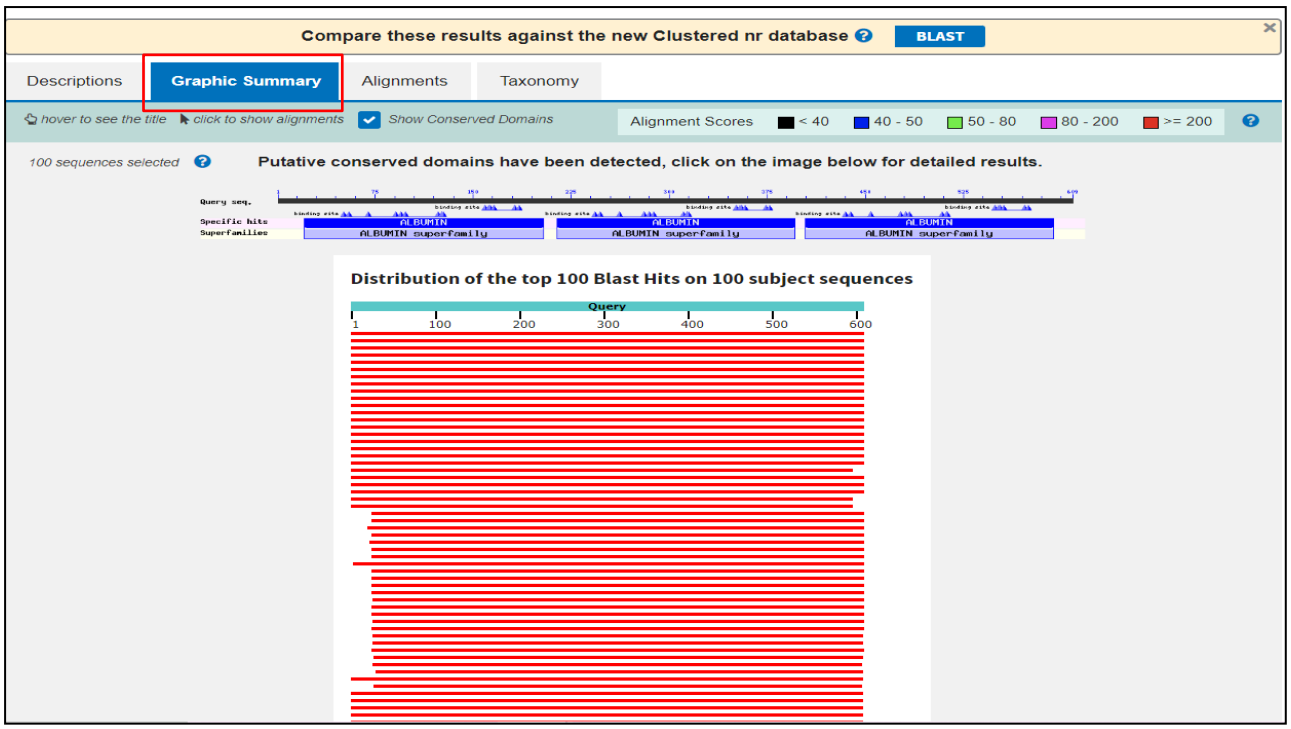


Figure 11: Result for Graphic Summary Section

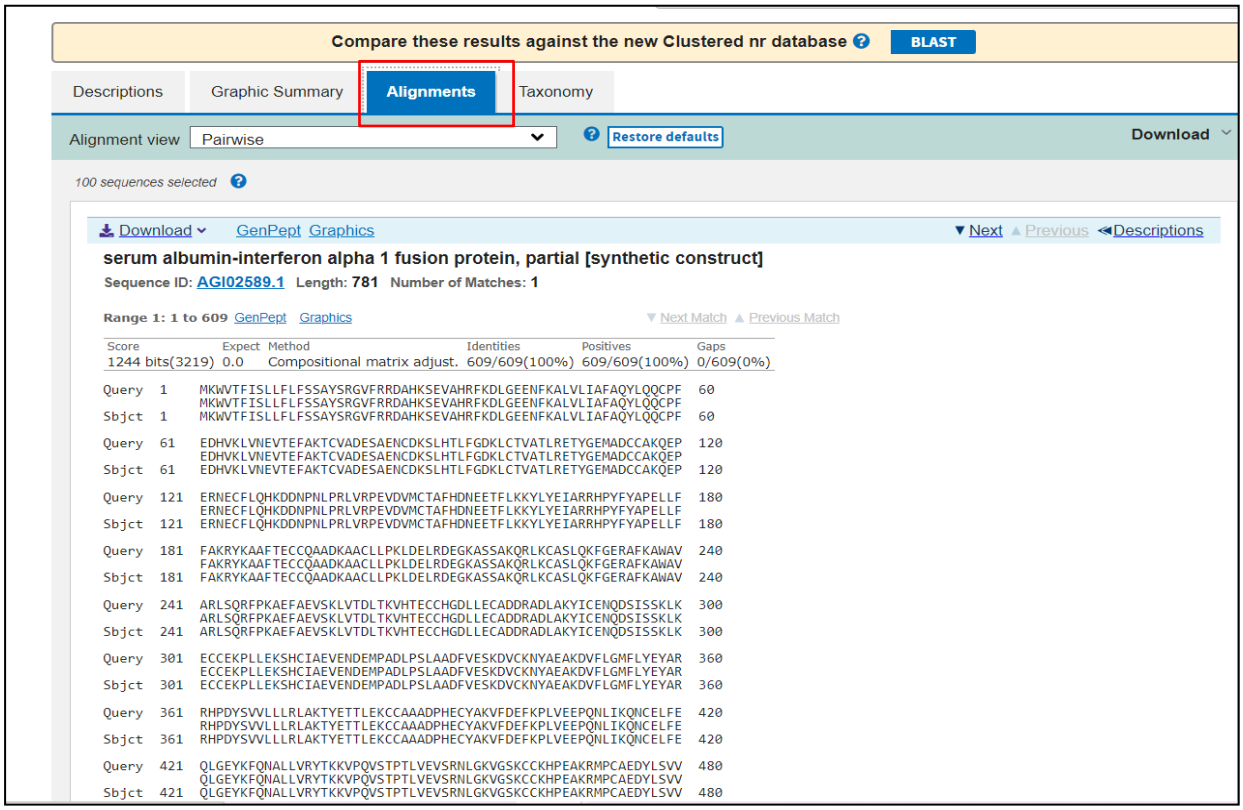
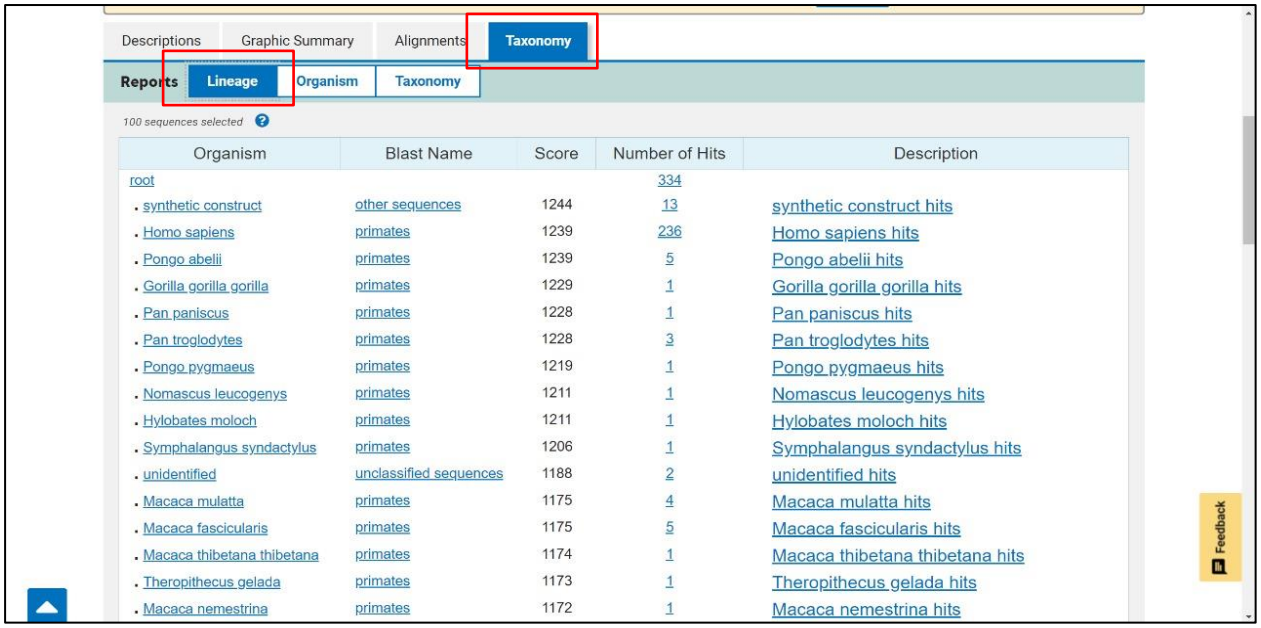


Figure 12: Result for Alignment Section



Descriptions | Graphic Summary | Alignments | Taxonomy

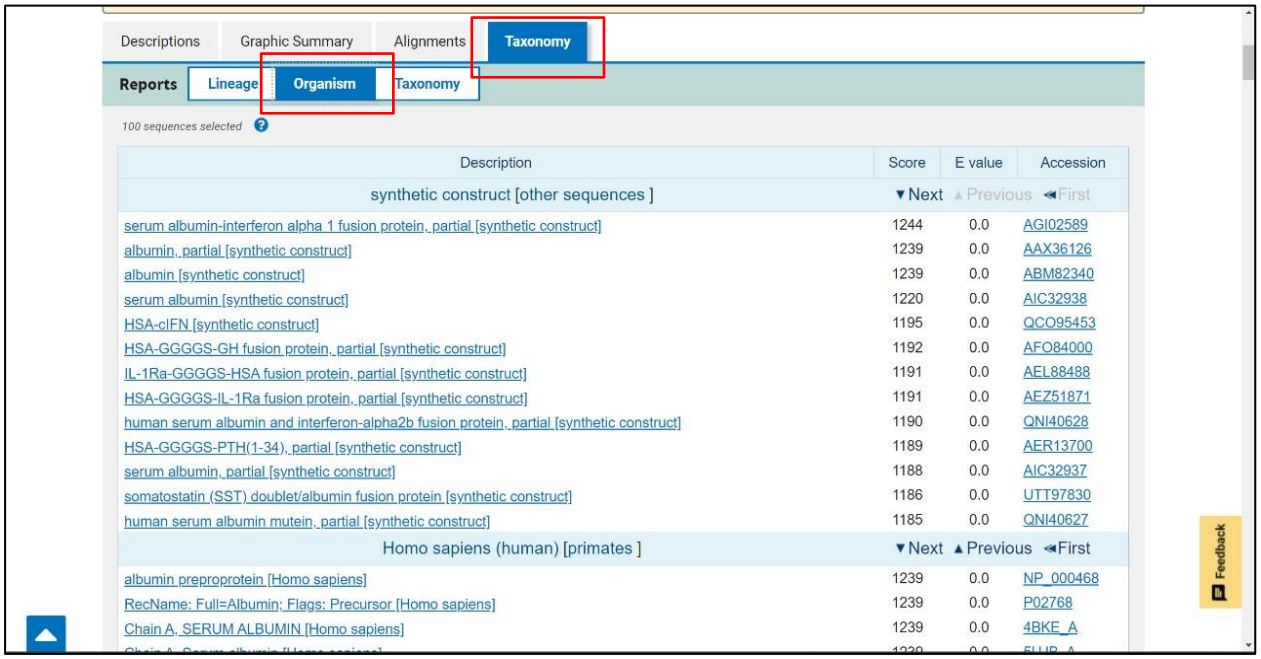
Reports | Lineage | Organism | Taxonomy

100 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
root			334	
synthetic construct	other sequences	1244	13	synthetic construct hits
Homo sapiens	primates	1239	236	Homo sapiens hits
Pongo abelii	primates	1239	5	Pongo abelii hits
Gorilla gorilla gorilla	primates	1229	1	Gorilla gorilla gorilla hits
Pan paniscus	primates	1228	1	Pan paniscus hits
Pan troglodytes	primates	1228	3	Pan troglodytes hits
Pongo pygmaeus	primates	1219	1	Pongo pygmaeus hits
Nomascus leucogenys	primates	1211	1	Nomascus leucogenys hits
Hylobates moloch	primates	1211	1	Hylobates moloch hits
Symphalangus syndactylus	primates	1206	1	Symphalangus syndactylus hits
unidentified	unclassified sequences	1188	2	unidentified hits
Macaca mulatta	primates	1175	4	Macaca mulatta hits
Macaca fascicularis	primates	1175	5	Macaca fascicularis hits
Macaca thibetana thibetana	primates	1174	1	Macaca thibetana thibetana hits
Theropithecus gelada	primates	1173	1	Theropithecus gelada hits
Macaca nemestrina	primates	1172	1	Macaca nemestrina hits

Feedback

Figure 13: Result for Taxonomy Section based on Lineage



Descriptions | Graphic Summary | Alignments | Taxonomy

Reports | Lineage | Organism | Taxonomy

100 sequences selected

Description	Score	E value	Accession
synthetic construct [other sequences]	▼Next	▲Previous	◀◀First
serum albumin-interferon alpha 1 fusion protein, partial [synthetic construct]	1244	0.0	AGI02589
albumin, partial [synthetic construct]	1239	0.0	AAX36126
albumin [synthetic construct]	1239	0.0	ABM82340
serum albumin [synthetic construct]	1220	0.0	AIC32938
HSA-cIFN [synthetic construct]	1195	0.0	QCO95453
HSA-GGGGS-GH fusion protein, partial [synthetic construct]	1192	0.0	AFO84000
IL-1Ra-GGGGS-HSA fusion protein, partial [synthetic construct]	1191	0.0	AEL88488
HSA-GGGGS-IL-1Ra fusion protein, partial [synthetic construct]	1191	0.0	AEZ51871
human serum albumin and interferon-alpha2b fusion protein, partial [synthetic construct]	1190	0.0	QNI40628
HSA-GGGGS-PTH(1-34), partial [synthetic construct]	1189	0.0	AER13700
serum albumin, partial [synthetic construct]	1188	0.0	AIC32937
somatostatin (SST) doublet/albumin fusion protein [synthetic construct]	1186	0.0	UTT97830
human serum albumin mutein, partial [synthetic construct]	1185	0.0	QNI40627
Homo sapiens (human) [primates]	▼Next	▲Previous	◀◀First
albumin preproprotein [Homo sapiens]	1239	0.0	NP_000468
RecName: Full=Albumin; Flags: Precursor [Homo sapiens]	1239	0.0	P02768
Chain A, SERUM ALBUMIN [Homo sapiens]	1239	0.0	4BKE_A

Feedback

Figure 13a: Result for Taxonomy Section based on Organism

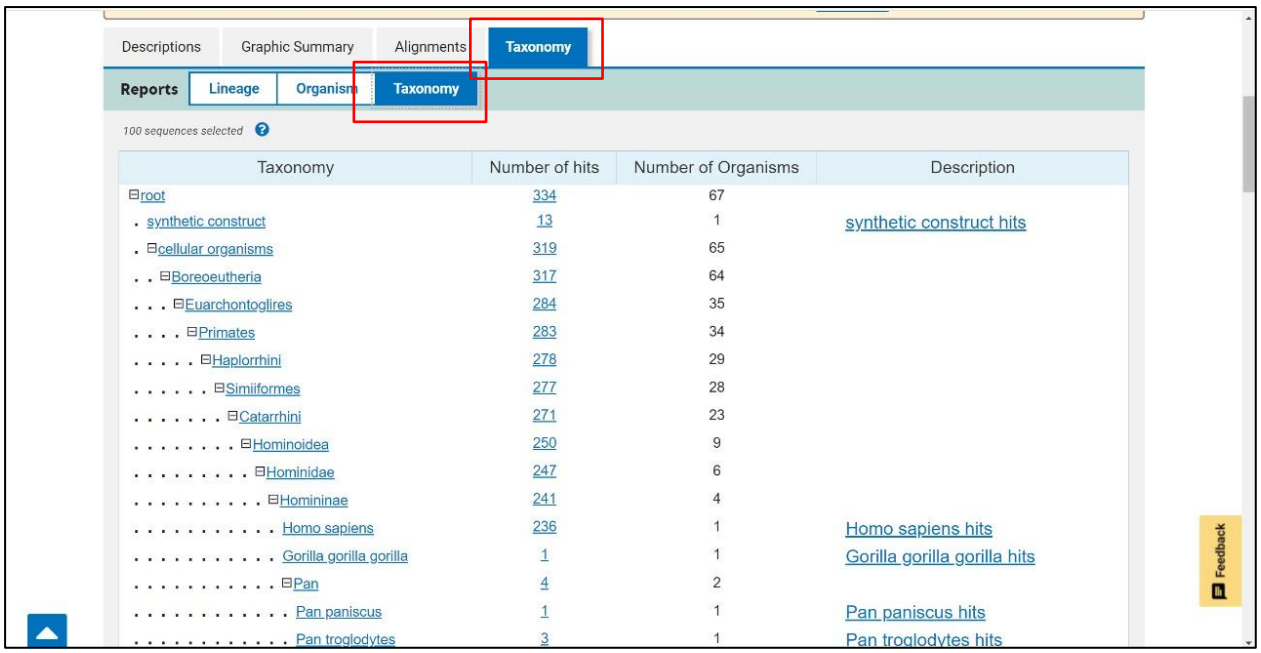


Figure 13b: Result for Taxonomy Section based on Taxonomy

RESULTS:

The Basic Local Alignment Search Tool (BLAST) was used to explore the protein sequences similar to the protein sequence of albumin (UniProt ID: P02768). The query sequence is found 100% identical to three sequence entries.

Sequence Title	Organism	Max Score	Total Score	E Value	Percentage Identity	Accession ID
serum albumin-interferon alpha 1 fusion protein	Synthetic construct	1244	1244	0.0	100.0%	AGI02589.1
albumin	Synthetic construct	1239	1239	0.0	100.0%	AAX36126.1
albumin preproprotein	*Homo Sapiens*	1239	1239	0.0	100.0%	NP_000468.1

CONCLUSION:

The protein sequences similar to the protein sequence of albumin (UniProt ID: P02768) were studied by exploring the Basic Local Alignment Search Tool (BLAST).

REFERENCES:

1. Xiong, J. (2006). *Essential Bioinformatics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511806087
2. S. Sugio, A. Kashima, S. Mochizuki, M. Noda, K. Kobayashi, Crystal structure of human serum albumin at 2.5 Å resolution, *Protein Engineering, Design and Selection*, Volume 12, Issue 6, June 1999, Pages 439–446, https://doi.org/10.1093/protein/12.6.439

3. He, X., Carter, D. Atomic structure and chemistry of human serum albumin. *Nature* 358, 209–215 (1992). https://doi.org/10.1038/358209a0

DATE: 01/11/2023

WEBLEM 6(B)
FASTA TOOL
(URL: https://www.ebi.ac.uk/Tools/sss/fasta/)

AIM:

To study protein sequence similarity by exploring FASTA tool for the query maltose (UniProt ID: P68187).

INTRODUCTION:

FASTA tool was originally developed for comparing protein sequences. FASTA is a text-based format for representing nucleotide or amino acid sequences. It's used in bioinformatics and biochemistry. FASTA is an abbreviation for "Fast-All". FASTA is a sequence alignment tool that takes nucleotide or protein sequences as input and compares it with existing databases. It was the first database similarity search tool developed, preceding the development of BLAST. The FASTA format allows for sequence names and comments to precede the sequences. Nucleotides or amino acids are represented using single-letter codes. For example, A => adenosine, C => cytidine, G => guanine, T => thymidine, and N => A/G/C/T (any). The original program was referred to as FASTP. It quickly became a popular tool for sequence alignment and database searching. The program has been continually updated and improved.

There are now different FASTA programs available, each used for different types of sequence searches:

1. **FASTA** compares a DNA query sequence against a database of DNA sequences or a protein query sequence against a database of protein sequences using the FASTA algorithm.
2. **SSEARCH** performs protein-protein or DNA-DNA comparisons using the SmithWaterman algorithm.
3. **GGSEARCH/GLSEARCH** works using a global alignment algorithm (GGSEARCH) or a combination of global and local alignment algorithms (GLSEARCH) to compare protein and nucleotide sequences.
4. **FASTX/FASTY** compares a DNA sequence and a database of protein sequences by translating the DNA sequence into three frames and allowing gaps and frameshifts.
5. **TFASTX/TFASTY** compares a protein sequence and a database of DNA sequences. The DNA sequence is translated in six frames – three in the forward direction and three in the reverse direction.
6. **FASTF/TFASTF** compares mixed peptide sequences against a protein (FASTF) or translated DNA (TFASTF) databases.
7. **FASTS/TFASTS** compares a set of short peptide fragments against the protein (FASTS) or translated DNA (TFASTS) databases.

1. **<u>How FASTA Works</u>**
   FASTA works by comparing a query sequence to a database of sequences to identify similar matches. The program uses a heuristic algorithm to quickly search the database and identify the most significant matches.

2. **<u>The working mechanism of FASTA is described in the following steps:</u>**
   Step 1: Identifying Regions
   The first step is identifying regions with high similarity by creating a lookup table for the query sequence. This step is also called hashing step. To create the lookup table, the query sequence is first broken down into smaller words known as k-tuples (ktup).

   Step 2: Re-Scoring
   In the second step, the ten best diagonals are rescored using suitable scoring matrices. For protein, BLOSUM50 or PAM matrix is used; for DNA sequences, the identity matrix is used. A subregion with the highest score is identified for each of the rescanned diagonal regions.

   Step 3: Joining Threshold
   Next, a score cutoff or the joining threshold is applied that excludes segments unlikely to be part of the final alignment. The library sequences are ranked based on their Initial scores.

   Step 4: Final Alignment
   Finally, the gapped alignment is refined to produce the final alignment. This is done by using the banded Smith-Waterman algorithm, which is a dynamic programming algorithm that calculates the optimal score (opt) for alignment.

<u>Maltose:</u>
Maltose-binding protein (MBP) is a part of the maltose/maltodextrin system of Escherichia coli, which is responsible for the uptake and efficient catabolism of maltodextrins. It is a complex regulatory and transport system involving many proteins and protein complexes. MBP has an approximate molecular mass of 42.5 kilodaltons.

<u>METHODOLOGY:</u>

1. The protein FASTA (canonical) sequence for the desired protein for the query of ‘Maltose’ (UniProt ID: P68187) was retrieved from the UniProt Database.
2. Open the homepage of EBI – FASTA tool. Select the desired Protein Database and paste the retrieved FASTA (canonical) sequence of Maltose (UniProt ID: P68187) in the query box of the EBI – FASTA tool.
3. Set the desired parameters and select the ‘SUBMIT’ option to submit the query to the tool.
4. The results were shown in different tabs, namely, Submission Information, Tool Output, Graphic Output, Functional Forecasts, and Summary Table.
5. Interpret the results obtained.

OBSERVATIONS:

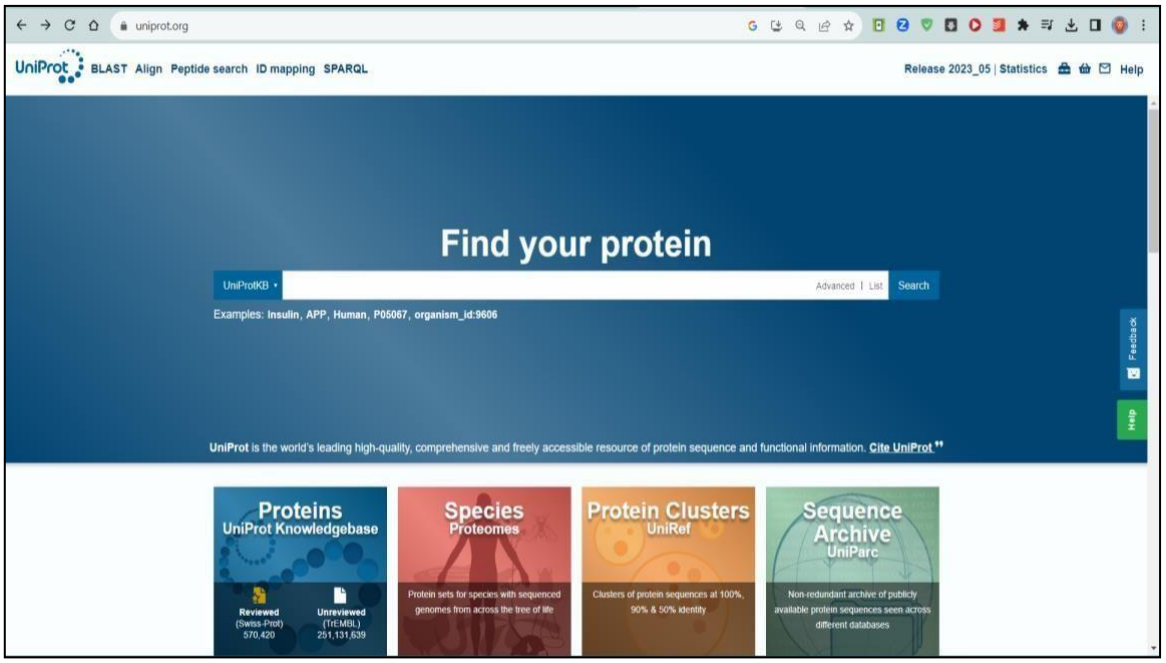


Figure 1: Homepage of the UniProt Database

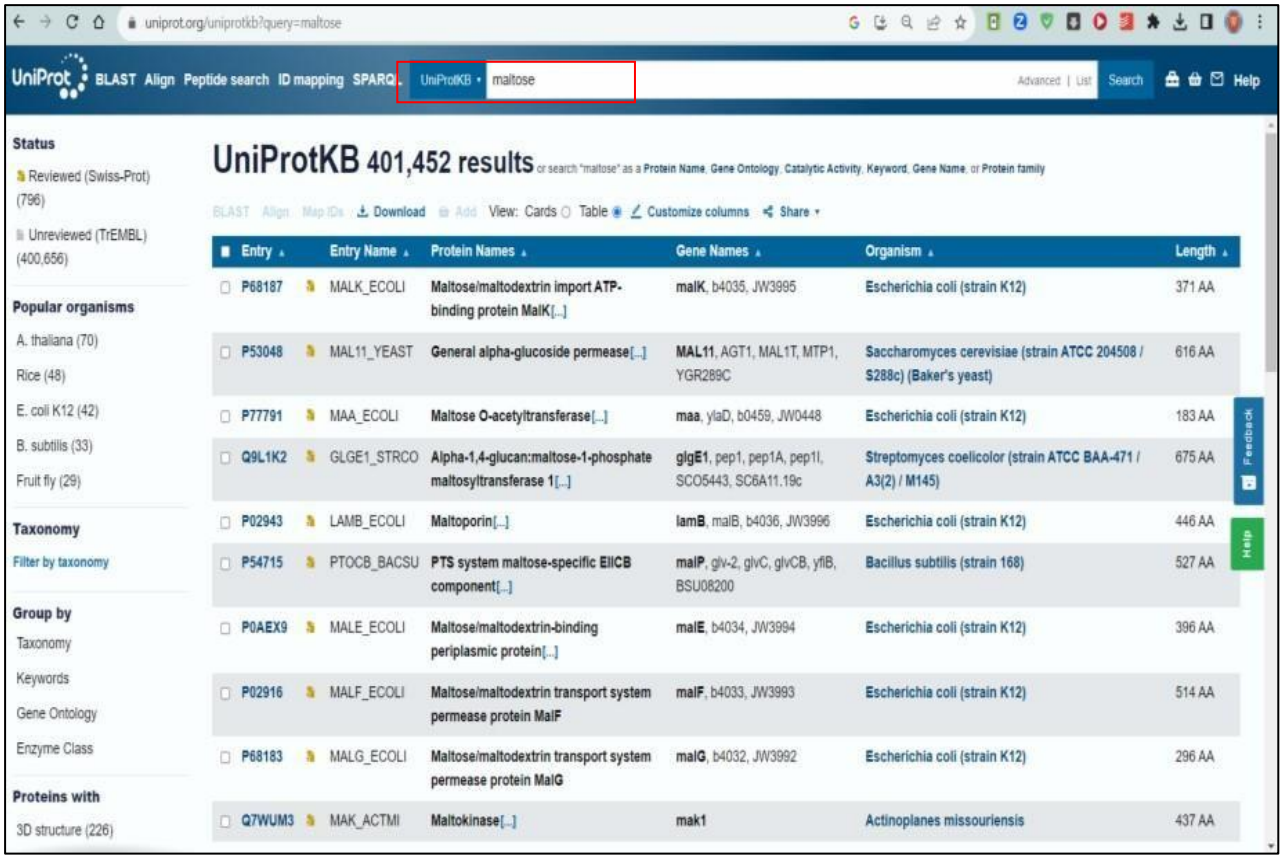


Figure 2: Searching for query maltose protein.

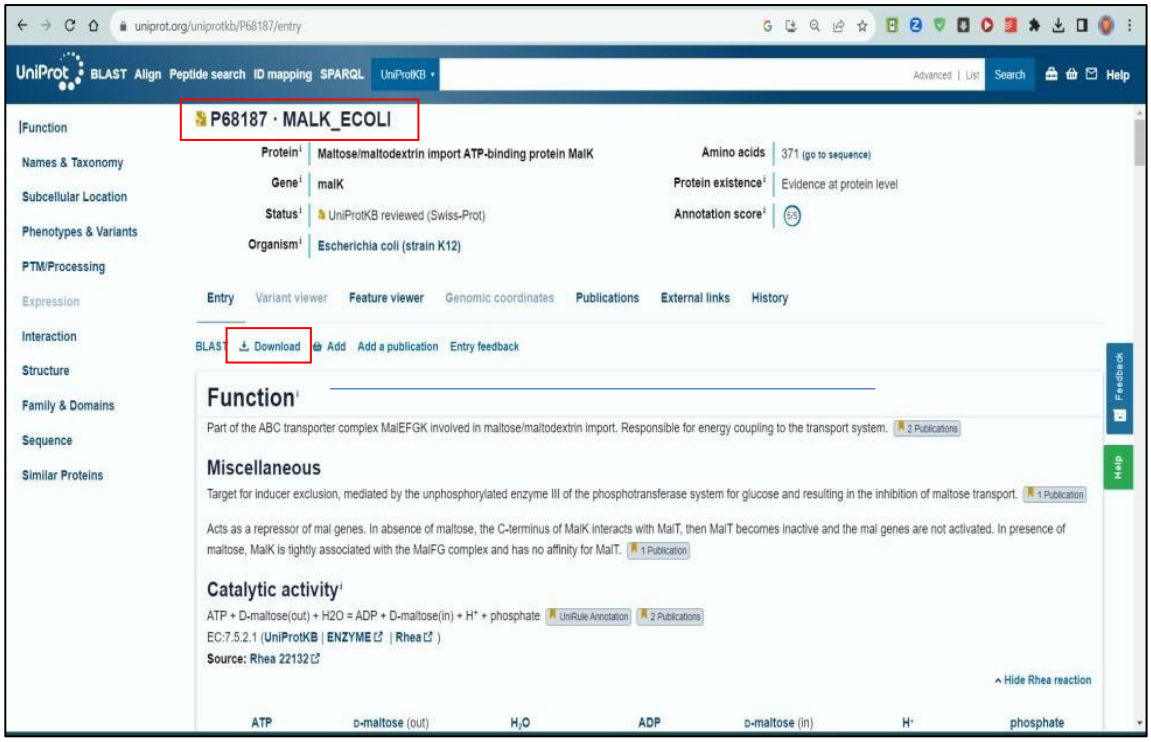


Figure 3: 'Download' option for retrieving the FASTA sequence of the protein

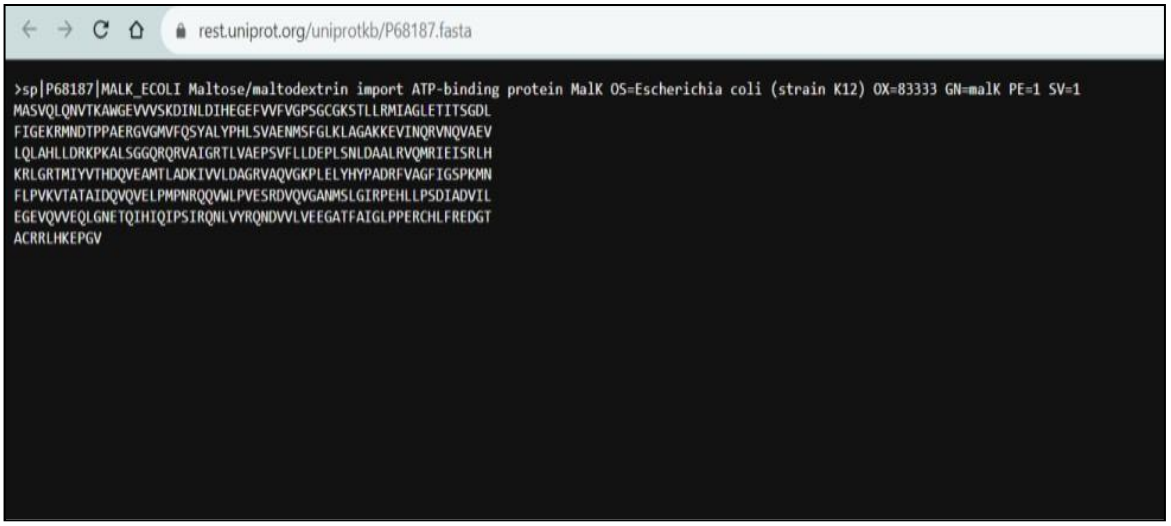


Figure 4: FASTA sequence of maltose protein.

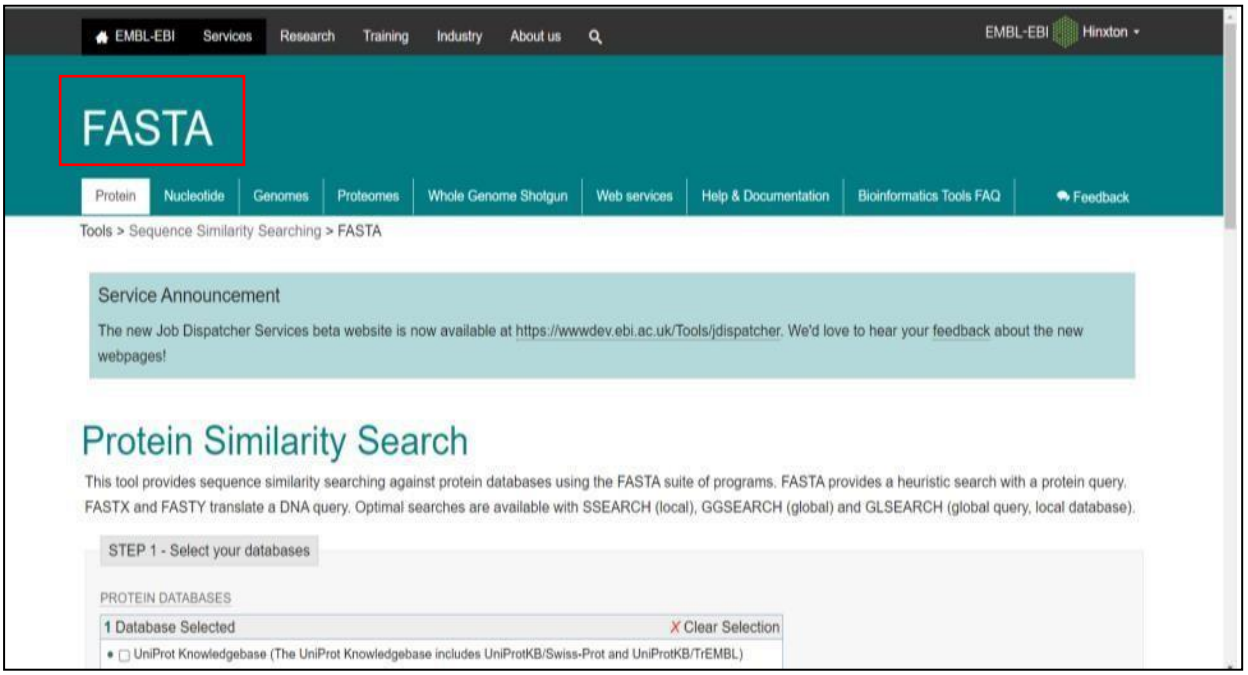


Figure 5: Homepage of FASTA tool.

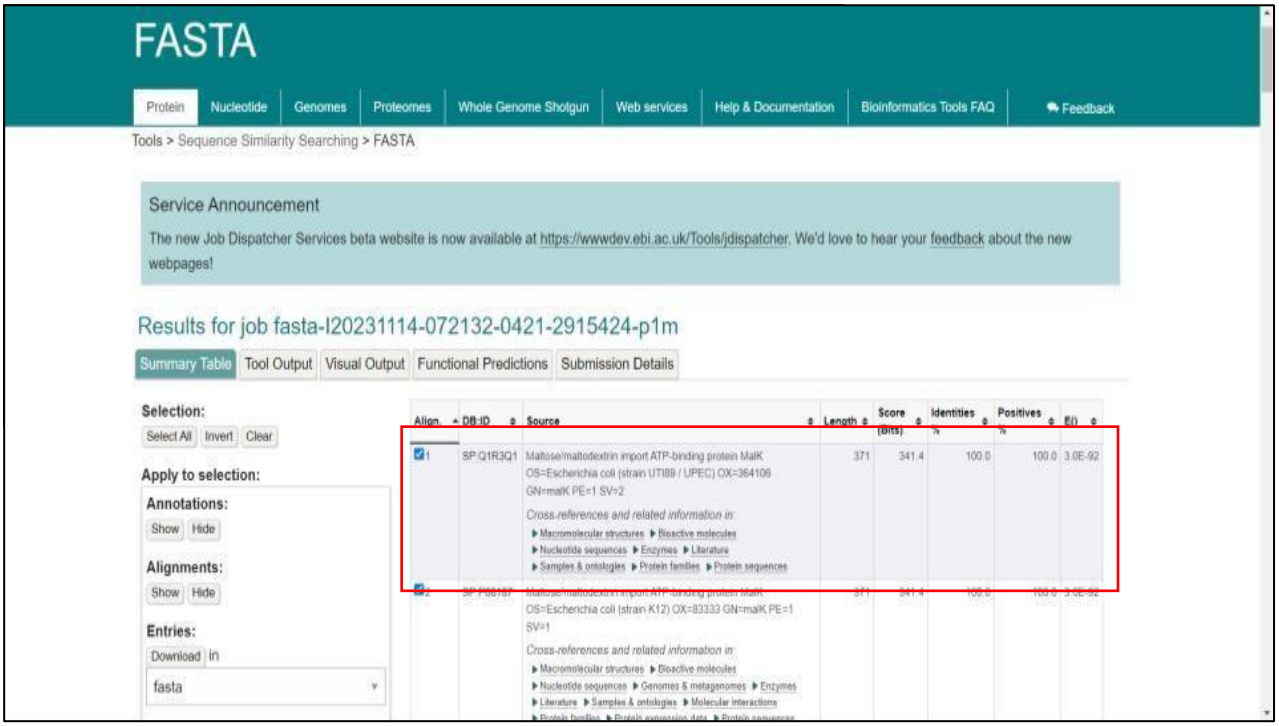


Figure 6: Searching sequence protein in FASTA tool.

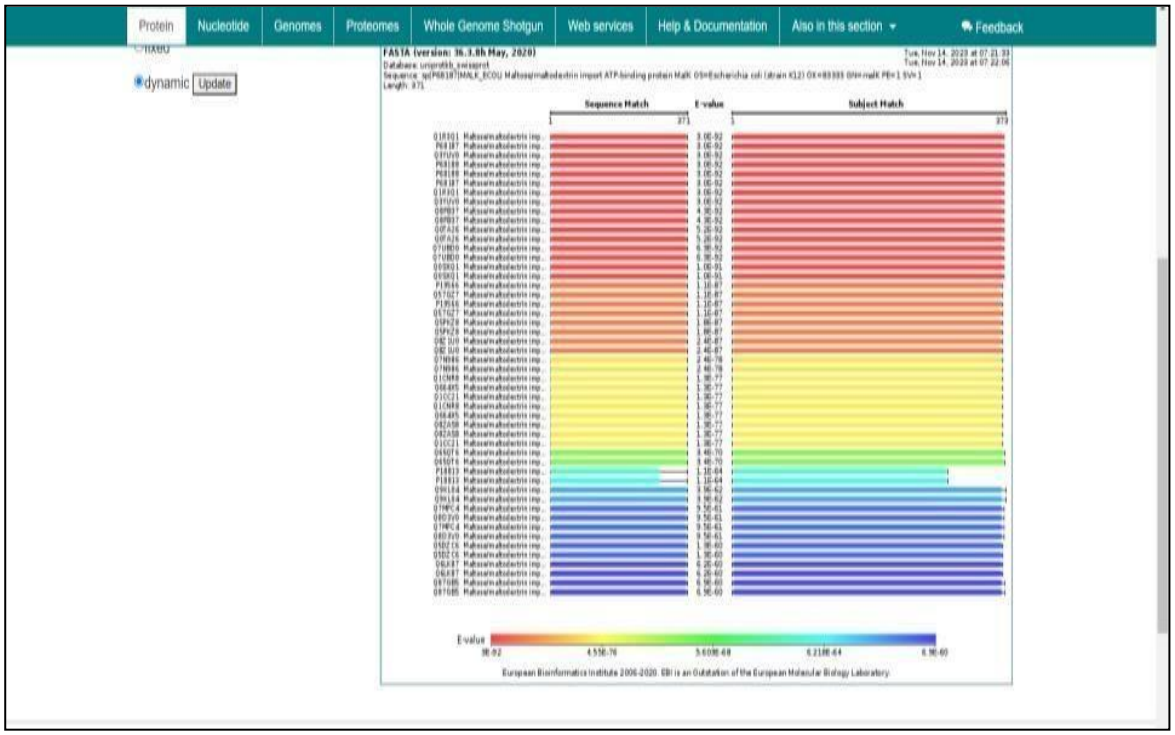


Figure 7: Visual output of maltose protein sequence.

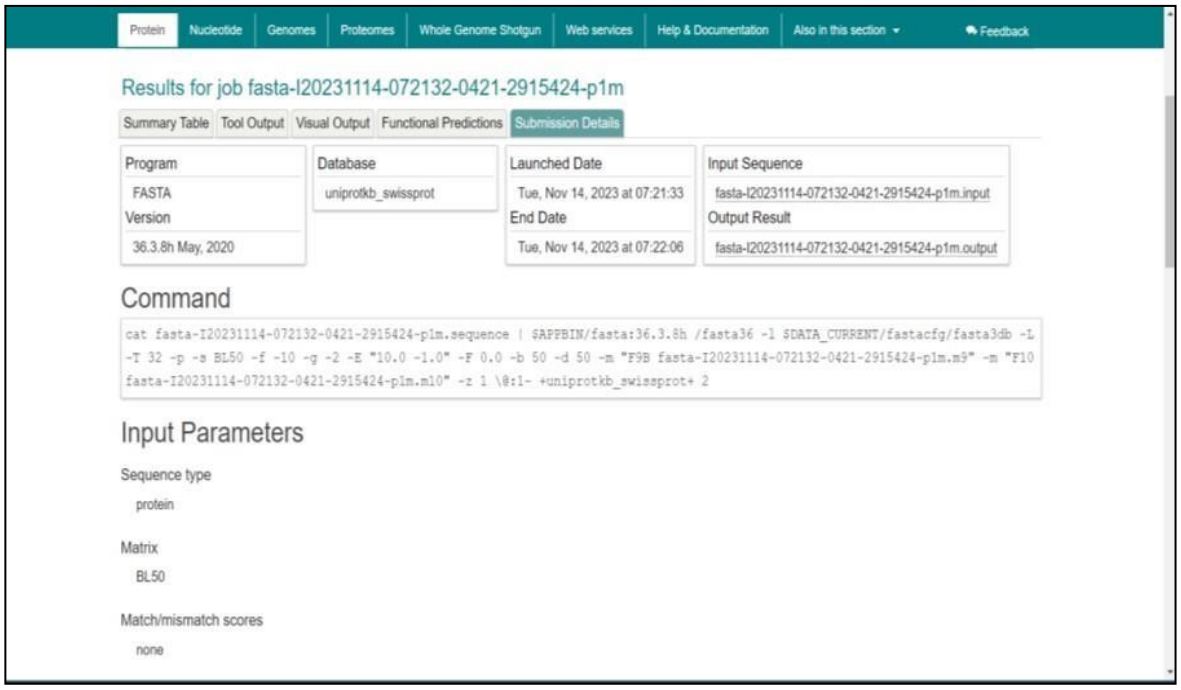


Figure 8: Submission details of maltose protein on FASTA tool.

RESULTS:

The EBI – FASTA tool was used to explore the sequences similar to the sequence of maltose (UniProt ID: P02768). The query sequence is found 100% identities & 100% positives to maltose sequence entries found in two organisms, viz., *Escherichia coli* and *Shigella sonnei*, with E Value of 5.2e-98 and sequence length of 371.

CONCLUSION:

FASTA is a versatile bioinformatics tool primarily used for storing, searching and comparing biological sequence data. It's commonly employed for tasks like sequence alignment, similarity searches and database comparisons. Sequence similarity was searched and studiedfor the Query 'Maltose' (UniProt ID: P68187) using the FASTA program.

REFERENCES:

1. Kryukov K, Ueda MT, Nakagawa S, Imanishi T (July 2020). "Sequence Compression Benchmark (SCB) database—A comprehensive evaluation of reference-free compressors for FASTA-formatted sequences". GigaScience. 9 (7): giaa072. https://doi.org/10.1093/gigascience/giaa072
2. Andrew Lloyd, Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (Methods of Biochemical Analysis, 43), Briefings in Bioinformatics, Volume 2, Issue 4, December 2001, Pages 407–408, https://doi.org/10.1093/bib/2.4.407
3. Pratas D, Hosseini M, Pinho A (2017). "Cryfa: a tool to compact and encrypt FASTA files". 11th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB). Advances in Intelligent Systems and Computing. Vol. 616. Springer. Pp. 305–312. Doi:10.1007/978-3-319-60816-7_37. https://link.springer.com/book/10.1007/978-3-319-60816-7

DATE: 01/11/2023

WEBLEM 6(C)
PROTEIN- SPECIFIC ITERATED BLAST (PSI BLAST)
(URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi)

AIM:

To explore the PSI BLAST tool to search putative homologs for query "Leucine" (UniProt ID: Q8IX15).

INTRODUCTION:

PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) derives a position-specific scoring matrix (PSSM) or profile from the multiple sequence alignment of sequences detected above a given score threshold using protein–protein BLAST. This PSSM is used to further search the database for new matches and is updated for subsequent iterations with these newly detected sequences. Thus, PSI-BLAST provides a means of detecting distant relationships between proteins. BLAST (Basic Local Alignment Search Tool) is a sequence similarity search method, in which a query protein or nucleotide sequence is compared to nucleotide or protein sequences in a target database to identify regions of local alignment and report those alignments that score a given score threshold. Position-Specific Iterative (PSI)-BLAST is a protein sequence profile search method that builds off the alignments generated by a run of the BLASTp program. The first iteration of a PSI-BLAST search is identical to a run of BLASTp program. It then generates a multiple alignment of the highest scoring pairs of the BLASTp run above a certain preset score or *e*-value threshold and calculates a profile or a position-specific score matrix (PSSM) from the multiple alignment.

The PSSM captures the conservation pattern in alignment and stores it as a matrix of scores for each position in the alignment-highly conserved positions receive high scores and weakly conserved positions receive scores near zero. This profile is used in place of the original substitution matrix for a further search of the database to detect sequences that match the conservation pattern specified by the PSSM. The newly detected sequences from this second round of the search, which are above the specified score (*e*-value) threshold is again added to alignment the profile is refined for another round of searching. This process is iteratively continued until desired or until convergence, i.e., the state where no new sequences are detected above the defined threshold. The iterative profile generation process makes PSI-BLAST far more capable of detecting distant sequence similarities than a single query alone in BLASTp, because it combines the underlying conservation information from a range of related sequence into a single score matrix. In the evolution, three-dimensional (3D) structures of proteins may be conserved even after considerable erosion of their sequence similarity. PSI-BLAST has been demonstrated to be useful in detecting such relationships via sequence searches, which were previously only detected through direct comparison of the 3D structures. Here, we discuss practical aspects of using PSI-BLST and provide a tutorial on how to uncover distant relationships between proteins and use them to reach biological meaningful conclusions.

Significance:

1. PSI-BLAST is most conveniently used on the internet with the help of the graphical user interface provided by the PSI-BLAST search page on National Centre for Biotechnology Information (NCBI).
2. The PSI-BLAST page may be customized by the user in terms of automated or semiautomated or "two-page formatting" and other parameters modified as desired. This page can then be saved as permanent internet bookmark for repeated use on futureoccasions.
3. As a rule of the thumb, beginners are advised to use the profile-inclusion threshold of expect (*e*)-value = 0.005 for their analysis. However, a user familiar with globular domains and compositional bias may use the inclusion threshold of 0.01 for inclusion in the profile, if a sequence does not have any major compositionally biased segments.
4. A pair of protein sequences can either be homologous (sharing a common evolutionary ancestor) or nonhomologous (evolutionarily unrelated).
   a. It should be noted that PSI-BLAST does not offer a direct binary decision on whether two sequences are related or not. However, the *e*-value obtained for a PSI-BLAST alignment can be used as a guide for this purpose.
5. As a heuristic it may be assumed that any compositionally unbiased query, encompassing a globular domain in a protein, giving a hit with *e*-value = <0.01 is likelyto be an indication of a homologous relationship. However, a user must carefully evaluate such alignments case-by-case because there can occasionally be false- positives.
6. A user may set the number of alignments and hits view as at least 1000 if searching the nonredundant (nr) database of NCBI, because of the large number hits obtained due tothe current size of the database. PSI-BLAST may also be downloaded and run as a standalone program for Windows or UNIX-type operating systems.
   a. However, in this case the various parameters need to be specified using the set of command-line flags for the program. An advantage of using the standalone version is the ability to use alignments as queries to generate a starting PSSM or saving and reusing the profile generated by a run of PSI-BLAST.

Leucine:

Leucine (symbol **Leu** or **L**) is essential amino acid that is used in the biosynthesis of proteins. Leucine is an α-amino acid, meaning it contains an α-amino group (which is in the protonated $-NH_3^+$ form under biological conditions), an α-carboxylic acid group (which is in the deprotonated $-COO^-$ form under biological conditions), and a side chain isobutyl group, making it a non-polar aliphatic amino acid. It is essential in humans, meaning the body cannot synthesize it: it must be obtained from the diet. Human dietary sources are foods that contain protein, such as meats, dairy products, soy products, and beans and other legumes. It is encoded by the codons UUA, UUG, CUU, CUC, CUA, and CUG.

Like valine and isoleucine, leucine is a branched-chain amino acid. The primary metabolic end products of leucine metabolism are acetyl-CoA and acetoacetate; consequently, it is one of the two exclusively ketogenic amino acids, with lysine being the other. It is the most important ketogenic amino acid in humans.

L-leucine is the L-enantiomer of leucine. It has a role as a plant metabolite, an Escherichia coli metabolite, a Saccharomyces cerevisiae metabolite, a human metabolite, an algal metabolite

and a mouse metabolite. It is a pyruvate family amino acid, a proteinogenic amino acid, a leucine and a L-alpha-amino acid. It is a conjugate base of a L-leucinium. It is a conjugate acid of a L-leucinate. It is an enantiomer of a D-leucine. It is a tautomer of a L-leucine zwitterion.

METHODOLOGY:

1. Go to the website of BLAST tool.
2. Click protein blast as protein is more conserved than nucleotide.
3. Go on UniProt portal.
4. Search for query 'Leucine'.
5. From shown results select UniProt ID: 'Q8IX15' entry.
6. Download the sequence in FASTA (Canonical) format.
7. Copy the sequence and paste under BLASTp suite.
8. Select Protein Data Bank (PDB) database under standard and program algorithm parameter as psi-blast with threshold 0.001.
9. Click BLAST to run the query.
10. Click Run to observe 2nd iterated and continue till 5 iterations.

OBSERVATIONS:

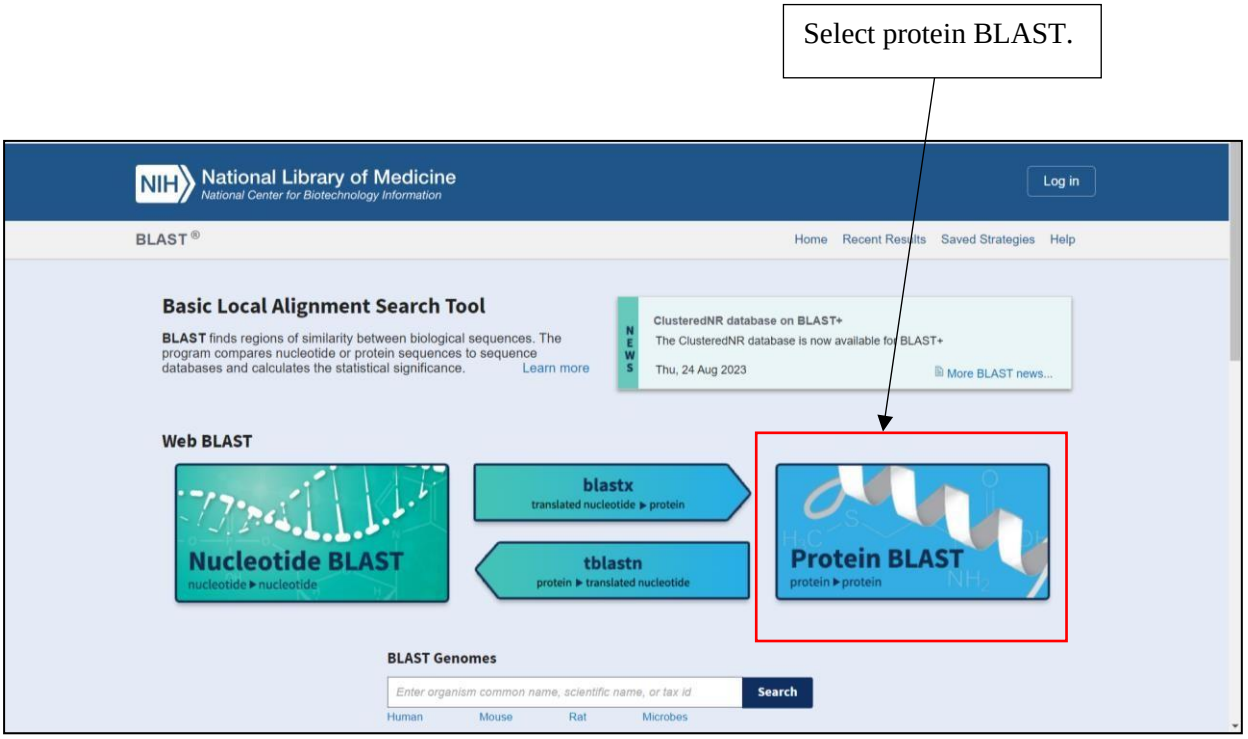


Figure 1: Homepage of BLAST

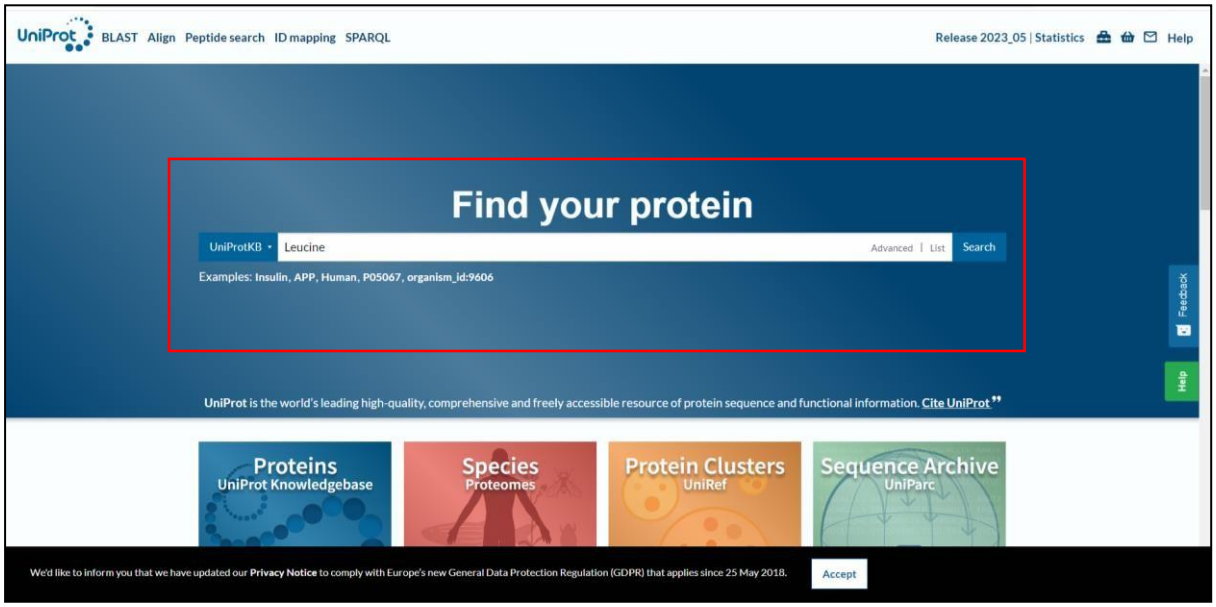


Figure 2: Query search in UniProt portal

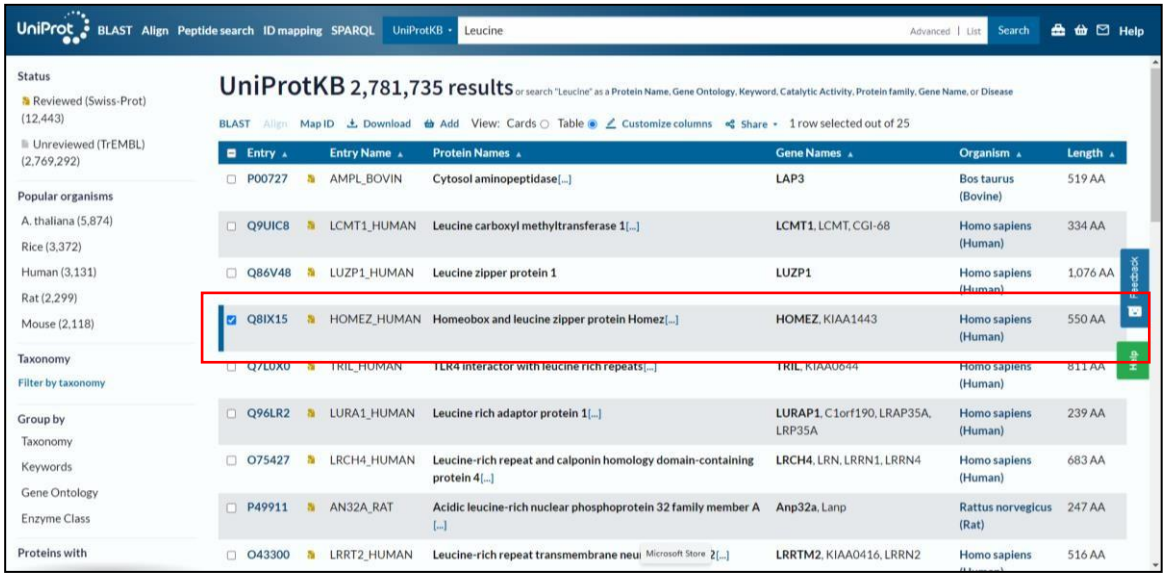


Figure 2a: Select desired organism

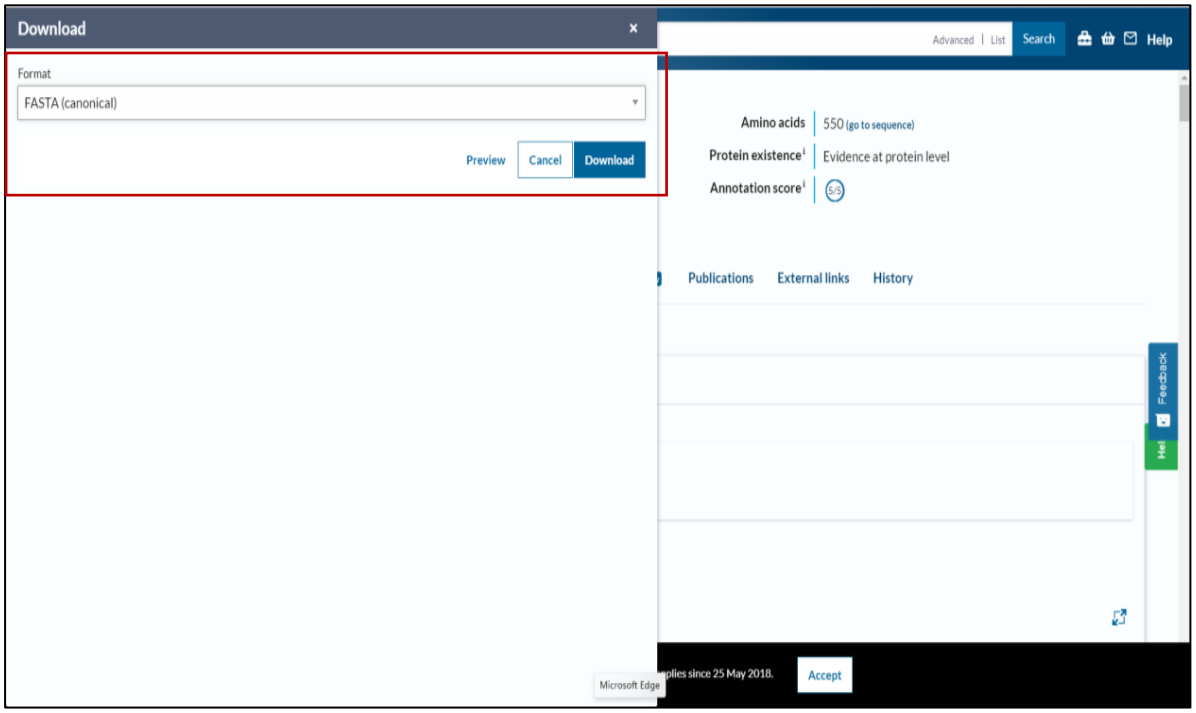


Figure 2b: Download sequence in FASTA (Canonical) format

```
>sp|Q8IX15|HOMEZ_HUMAN Homeobox and leucine zipper protein Homez OS=Homo sapiens OX=9606 GN=HOMEZ PE=1 SV=2
MVRGWEPPPGLDCAISEGHKSEGTMPPNKEASGLSSSPAGLICLPPISEELQLVWTQAAQ
TSELDSNEHLLKTFSYFPYPSLADIALLCLRYGLQMEKVKTWFMAQRLRCGISWSSEEIE
ETRARVVYRRDQLHFKSLLSFTHHAGRPPEEVPPPPVPAPEQVGIGIGPPTLSKPTQTKG
LKVEPEEPSQMPPLPQSHQKLKESLMTPGSGAFPYQSDFWQHLQSSGLSKEQAGRGPNQS
HGIGTASWNHSTTVPQPQARDKPPPIALIASSCKEESASSVTPSSSSTSSSFQVLANGAT
AASKPLQPLGCVPQSVSPSEQALPPHLEPAWPQGLRHNSVPGRVGPTEYLSPDMQRQRKT
KRKTKEQLAILKSFFLQCQWARREDYQKLEQITGLPRPEIIQWFGDTRYALKHGQLKWFR
DNAVPGAPSFQDPAIPTPPPSTRSLNERAETPPLPIPPPPPDIQPLERYWAAHQQLRETD
IPQLSQASRLSTQQVLDWFDSRLPQPAEVVVCLDEEEEEEEEELPEDDEEEEEEEEEDDD
DDDDDVIIQD
```

Figure 2c: Copying the sequence

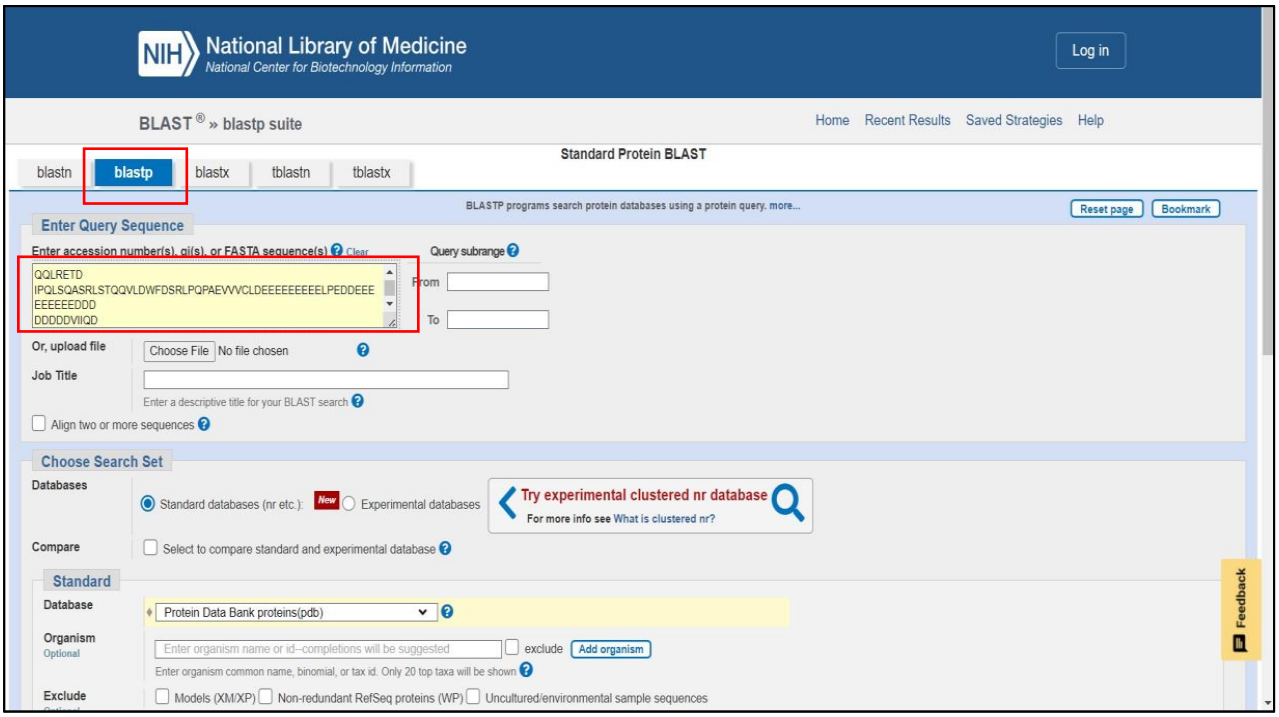


Figure 3: Pasting the sequence in BLASTp format

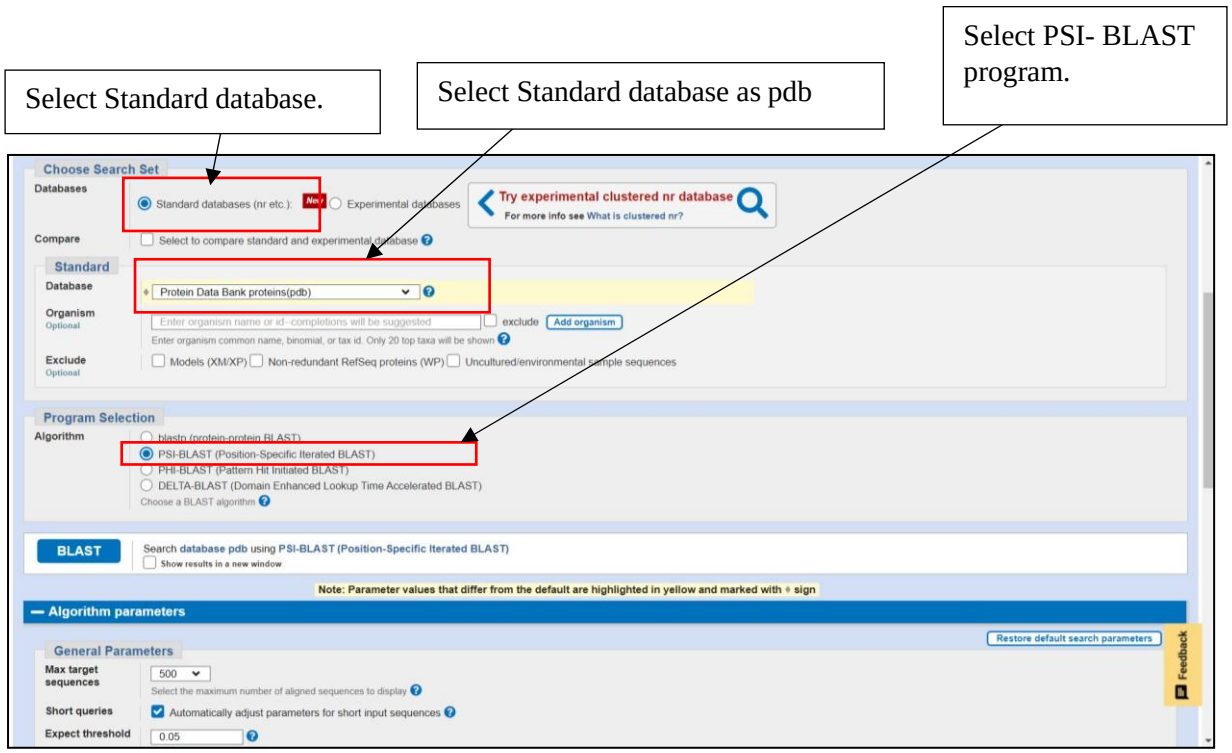


Figure 4: Selecting Standard database as pdb and program selection as PSI- BLAST

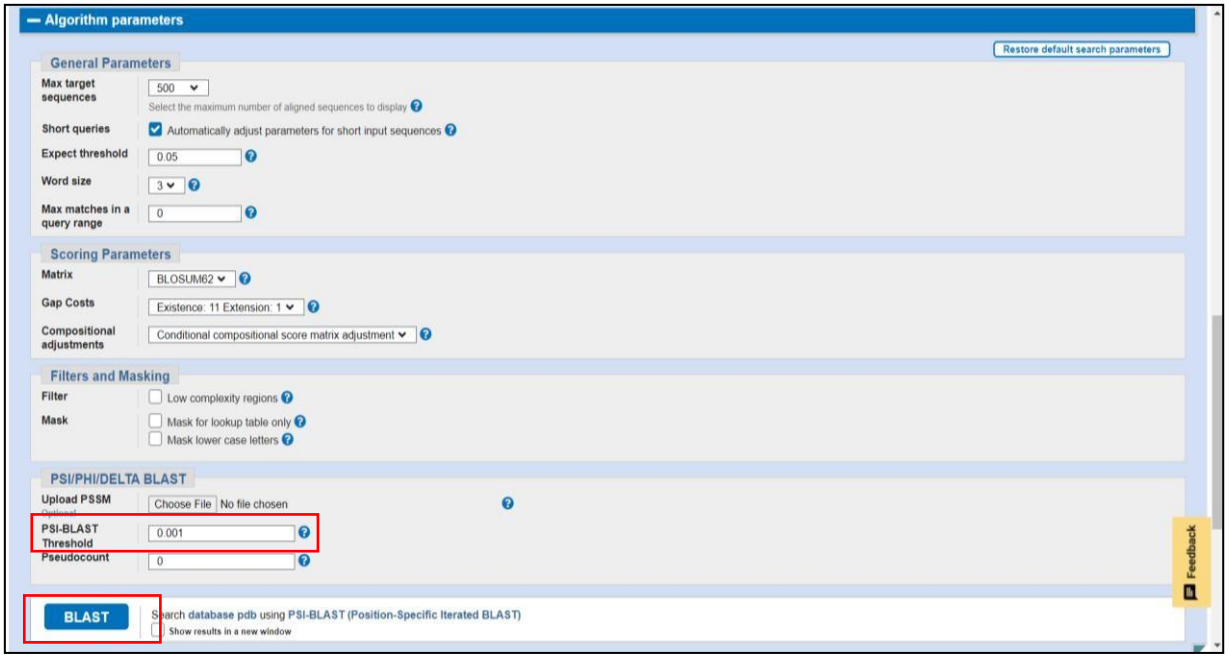


Figure 5: Keeping PSI-BLAST threshold as 0.001 and running PSI - BLAST

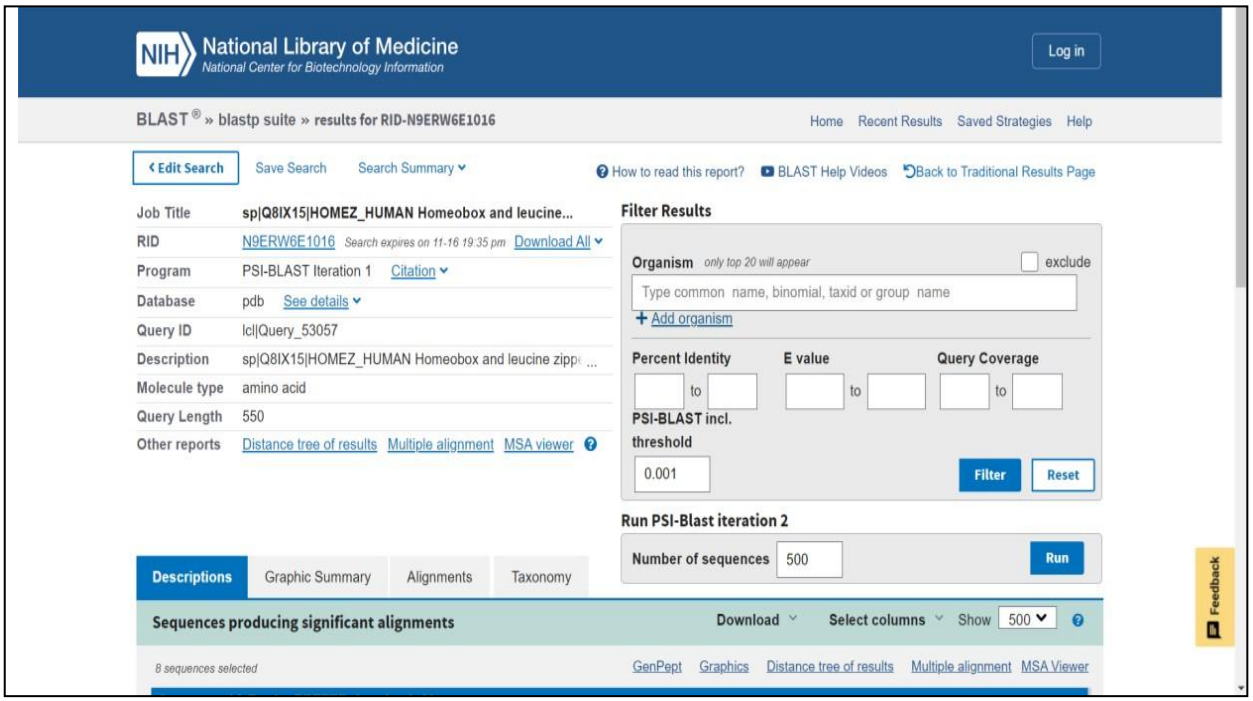


Figure 6: Result shown for UniProt ID: Q8IX15 in BLASTp

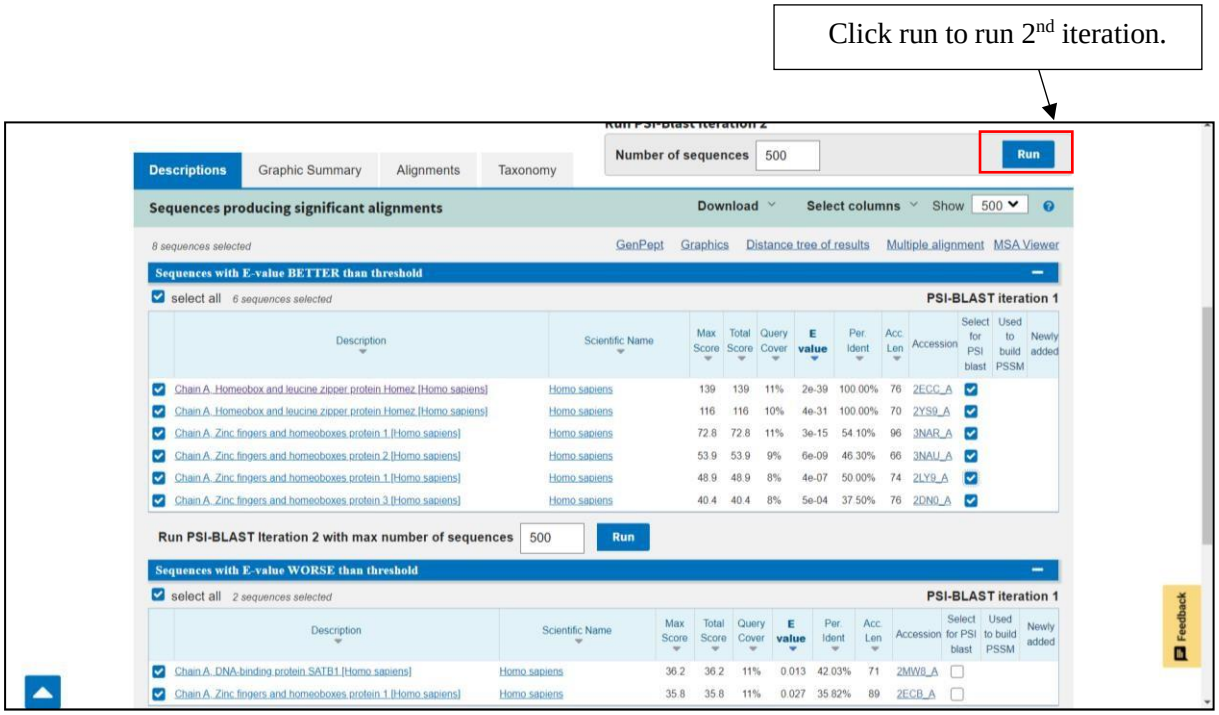


Figure 6a: Result shown for sequence with E- value better and worse than threshold

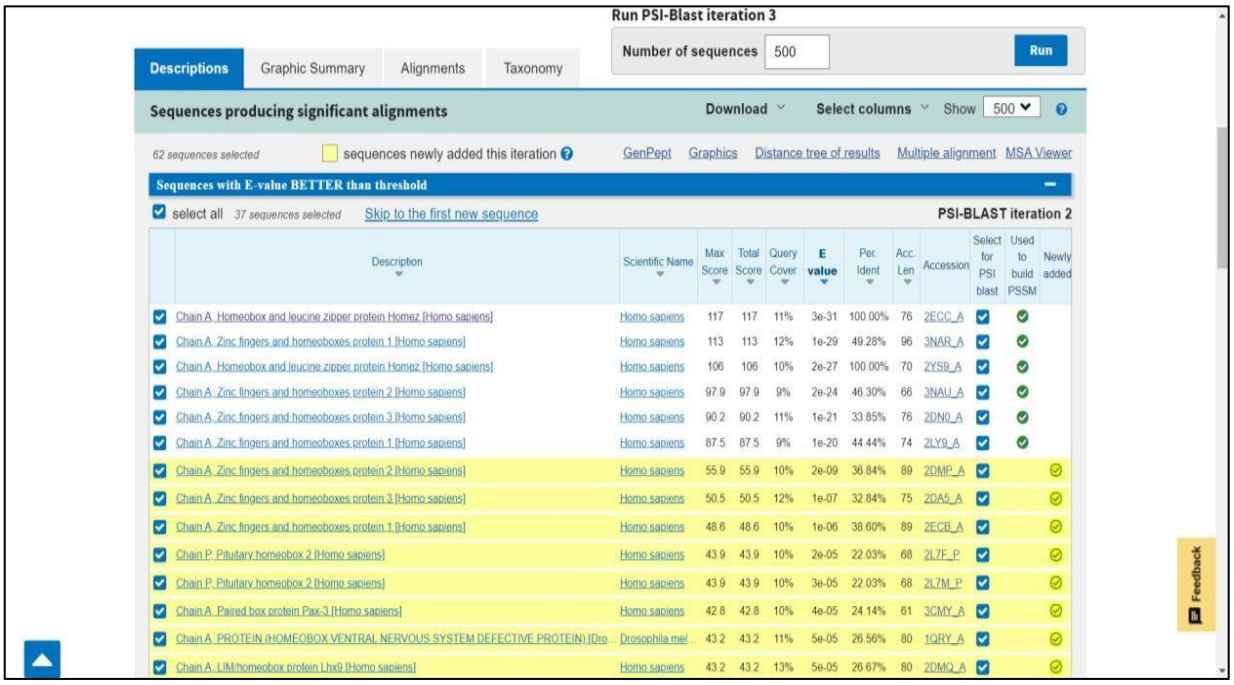


Figure 7: 2nd iterated result of UniProt ID: Q8IX15 organism

RESULTS:

PSI BLAST was explored using query 'Leucine' (Q8IX15) in order to get putative homologs. The first iteration showed 8 new putative sequences and the addition of new sequences was carried till 5th iteration, but then the process if halted as further iteration would drop the result accuracy and the iteration showed that new putative homologs are available for query 'Leucine'.

CONCLUSION:

PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) derives a position-specific scoring matrix (PSSM) or profile from the multiple sequence alignment of sequences detected above a given score threshold using protein–protein BLAST. This PSSM is used to further search the database for new matches and is updated for subsequent iterations with these newly detected sequences. Thus, PSI-BLAST provides a means of detecting distant relationships between proteins. PSI-BLAST (Position specific iterative – BLAST) algorithm program was used to view and explore best iterated results for query 'Leucine' (UniProt ID: Q8IX15).

REFERENCES:

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410.
2. PSI-BLAST. (n.d.). National Institutes of Health. https://www.ncbi.nlm.nih.gov/books/NBK2590/
3. Pruitt KD, Tatusova T, Ostell JM. McEntyre J, Ostell J, editors. The Reference Sequence (RefSeq) Project. National Library of Medicine (US), NCBI; Bethesda, MD: The NCBI Handbook. 2005 Chapter 18.
4. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. (1997 September 01) *Nucleic acids research* 25 (17) :3389-3402
5. Zhang, L., Li, F., Guo, Q., Duan, Y., Wang, W., Zhong, Y., Yang, Y., & Yin, Y. (2020). Leucine Supplementation: A Novel Strategy for Modulating Lipid Metabolism and Energy Homeostasis. *Nutrients*, *12*(5), 1299. https://doi.org/10.3390/nu12051299

DATE: 01/11/2023

WEBLEM 6(D)
PATTERN HIT INITIATED BLAST (PHI-BLAST) TOOL
(URL: https://blast.ncbi.nlm.nih.gov)

AIM:

To perform iterative blast for query 'Flavodoxin' protein (UniProt ID: P53554) by exploring Pattern Hit Initiated BLAST (PHI-BLAST) Tool.

INTRODUCTION:

Pattern Hit Initiated BLAST (PHI-BLAST) Tool, represents a variant of the BLAST algorithm employed for searching a protein database to identify other instances of a specific pattern occurring at least once within the input sequence. It facilitates the alignment and construction of the Position-Specific Scoring Matrix (PSSM) around a motif present in the query sequence. PHI-BLAST was developed by Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, and David J. Lipmann at the National Institutes of Health (NIH).

PHI-BLAST finds application in the analysis of various protein sequences, including CED4-like cell death regulators, HS90-type ATPase domains, archaeal tRNA nucleotidyltransferases, and archaeal proteins. It is utilized to identify protein sequences containing a specific pattern specified by the user and similar to the query sequence.

Compared to other BLAST tools, PHI-BLAST offers advantages such as increased speed and the ability for the user to express a rigid pattern occurrence requirement. This feature aids in reducing the number of hits that solely contain the pattern but lack true homology to the query sequence. However, PHI-BLAST may have a potential disadvantage in that it might be less sensitive than PSI-BLAST for detecting remote homologs. Additionally, the use of a specific pattern may restrict the search scope, potentially causing the omission of homologs lacking the specified pattern.

Flavodoxin:

Flavodoxins are small, soluble, electron-transfer proteins. Flavodoxins contains flavin mononucleotide as prosthetic group. The structure of flavodoxin is characterized by a five-stranded parallel beta sheet, surrounded by five alpha helices. They have been isolated from prokaryotes, cyanobacteria, and some eukaryotic algae. It functions in various metabolic processes, including photosynthesis, nitrogen and fatty acid metabolism. Flavodoxin is also involved in the detoxification of reactive oxygen species. The protein is reduced by flavodoxin reductase and transfers electrons to various redox enzymes. The semiquinone conformation of flavodoxin is stabilized by a hydrogen bond to the N-5 position of flavin, and a common tryptophan residue near the binding site aids in lowering SQ reactivity. The hydroquinone form is forced into a planar conformation, destabilizing it.

METHODOLOGY:

1. Open the homepage of UniProt database and search for the query 'Flavodoxin' protein.
2. Select any one entry from the results e.g., *Bacillus subtilis (strain 168)* **(**UniProt ID: P53554) and download its FASTA sequence in canonical format.
3. Open the homepage of BLAST and click on protein BLAST.
4. Paste the FASTA sequence in 'Enter query sequence' box and in program selection click on PHI-BLAST option.
5. Open the homepage of PROSITE database and search for the query 'Flavodoxin' protein.
6. Enter the FASTA sequence in 'Quick Scan mode of ScanProsite' box and scan it.
7. Copy the decoded pattern and paste it in the pattern in 'Enter a PHI pattern' box on PHI-BLAST portal and set the desired algorithm parameters.
8. Run the PHI-BLAST.
9. After each iteration, the new sequences are added to the results. These new sequences are highlighted using yellow color.
10. Run the PHI-BLAST iteration for 3-5 times, post which it starts generating garbage results, due to the decrease in sensitivity.
11. Interpret the results obtained.

OBSERVATIONS:

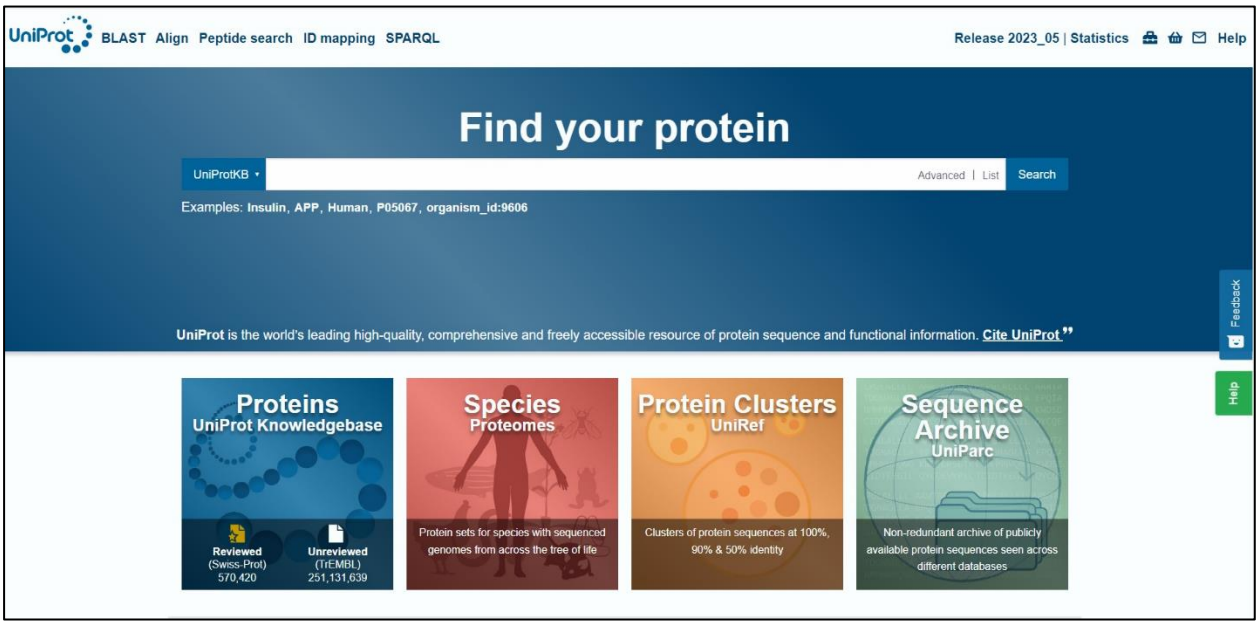


Figure 1: Homepage of the UniProt database

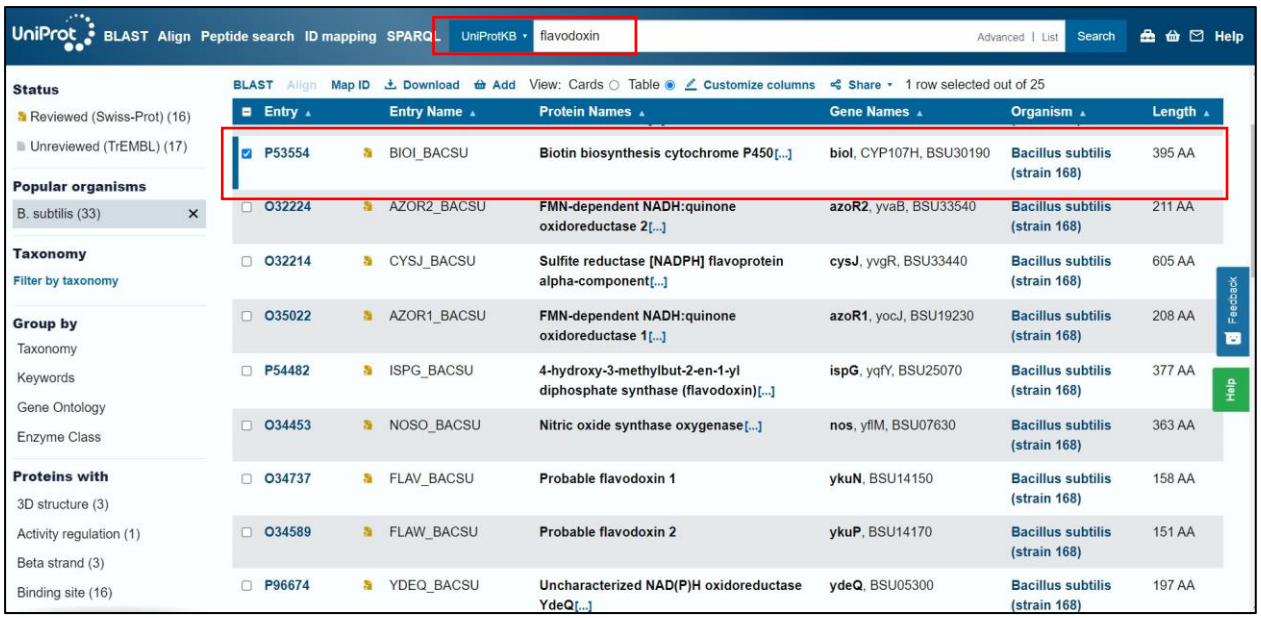


Figure 2: Query search for 'Flavodoxin' protein

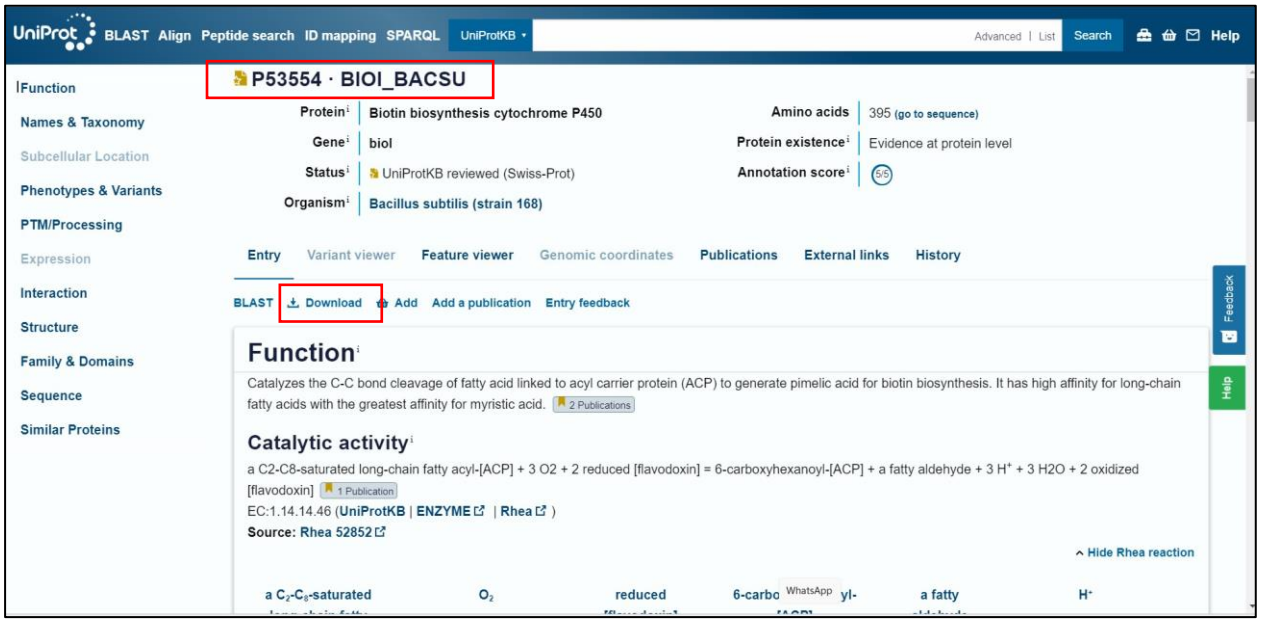


Figure 2a: Downloading the FASTA sequence for selected UniProt ID: P53554

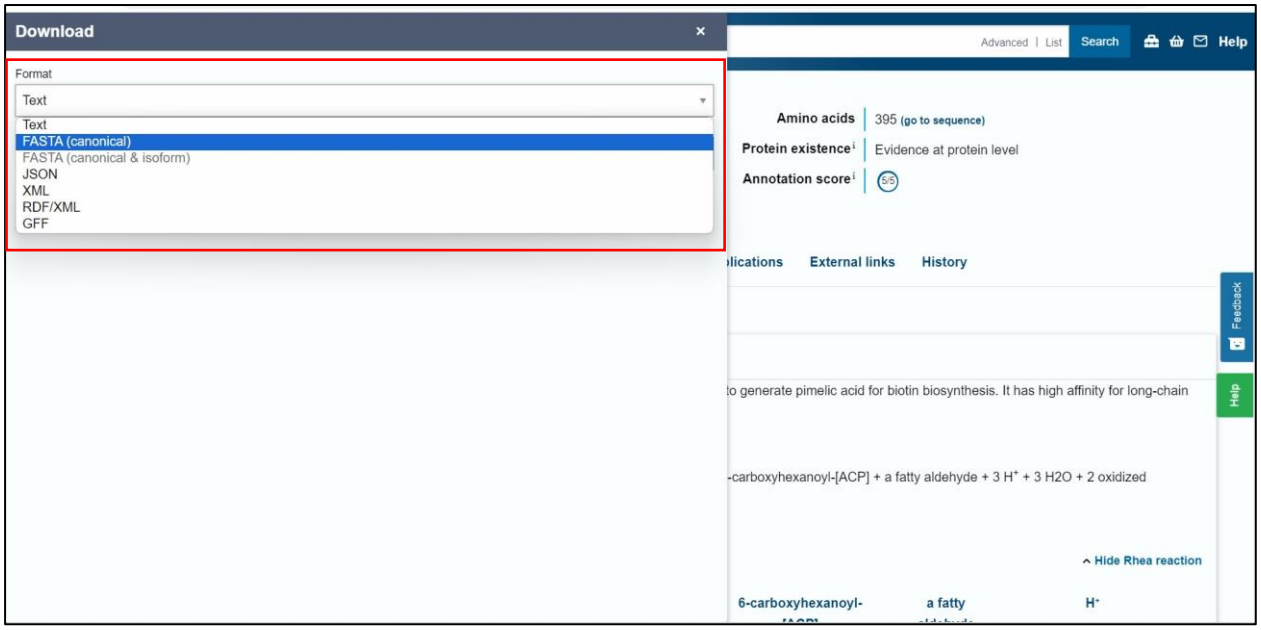


Figure 2b: Downloading the FASTA sequence in canonical format

```
>sp|P53554|BIOI_BACSU Biotin biosynthesis cytochrome P450 OS=Bacillus subtilis (strain 168) OX=224308 GN=bioI PE=1 SV=1
MTIASSTASSEFLKNPYSFYDTLRAVHPIYKGSFLKYPGWYVTGYEETAAILKDARFKVR
TPLPESSTKYQDLSHVQNQMMLFQNQPDHRRLRTLASGAFTPRTTESYQPYIIETVHHLL
DQVQGKKKMEVISDFAFPLASFVIANIIGVPEEDREQLKEWAASLIQTIDFTRSRKALTE
GNIMAVQAMAYFKELIQKRKRHPQQDMISMLLKGREKDKLTEEEAASTCILLAIAGHETT
VNLISNSVLCLLQHPEQLLKLRENPDLIGTAVEECLRYESPTQMTARVASEDIDICGVTI
RQGEQVYLLLGAANRDPSIFTNPDVFDITRSPNPHLSFGHGHHVCLGSSLARLEAQIAIN
TLLQRMPSLNLADFEWRYRPLFGFRALEELPVTFE
```

Figure 2c: View of the downloaded FASTA sequence

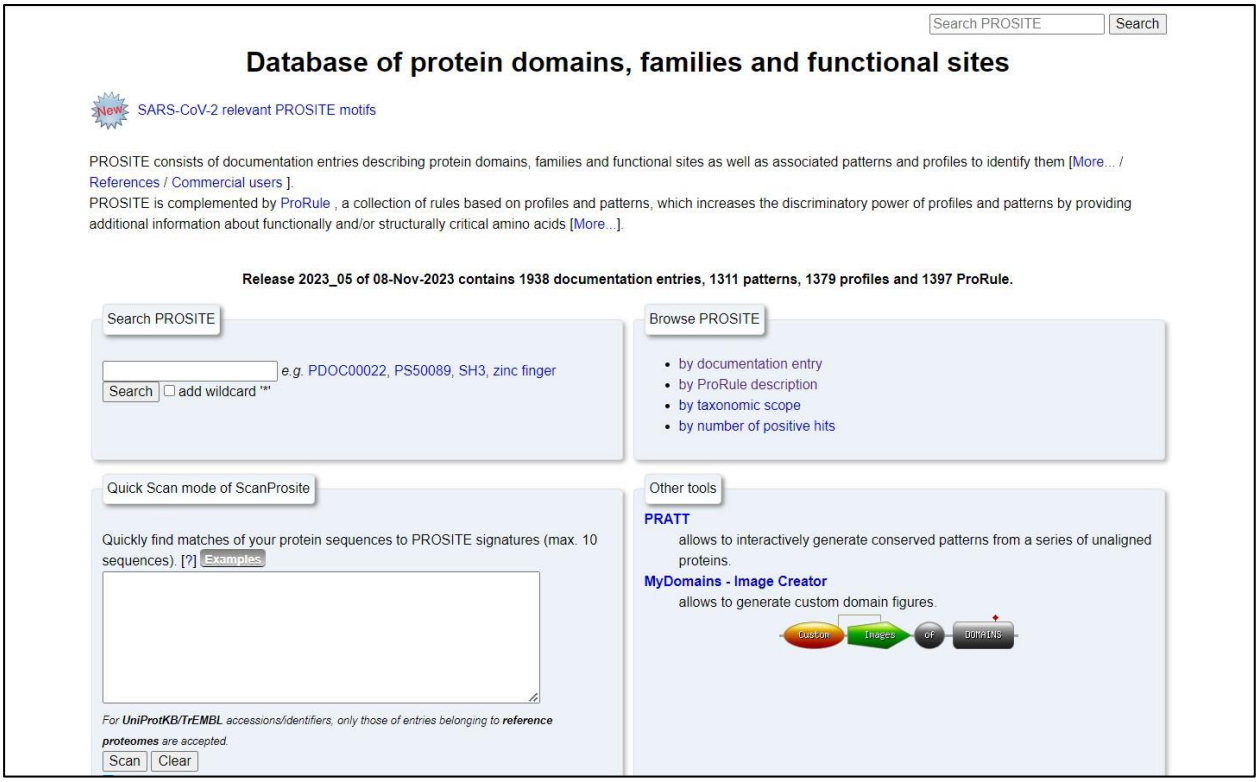


Figure 3: Homepage of PROSITE Database

Figure 3a: Paste the downloaded FASTA sequence for pattern

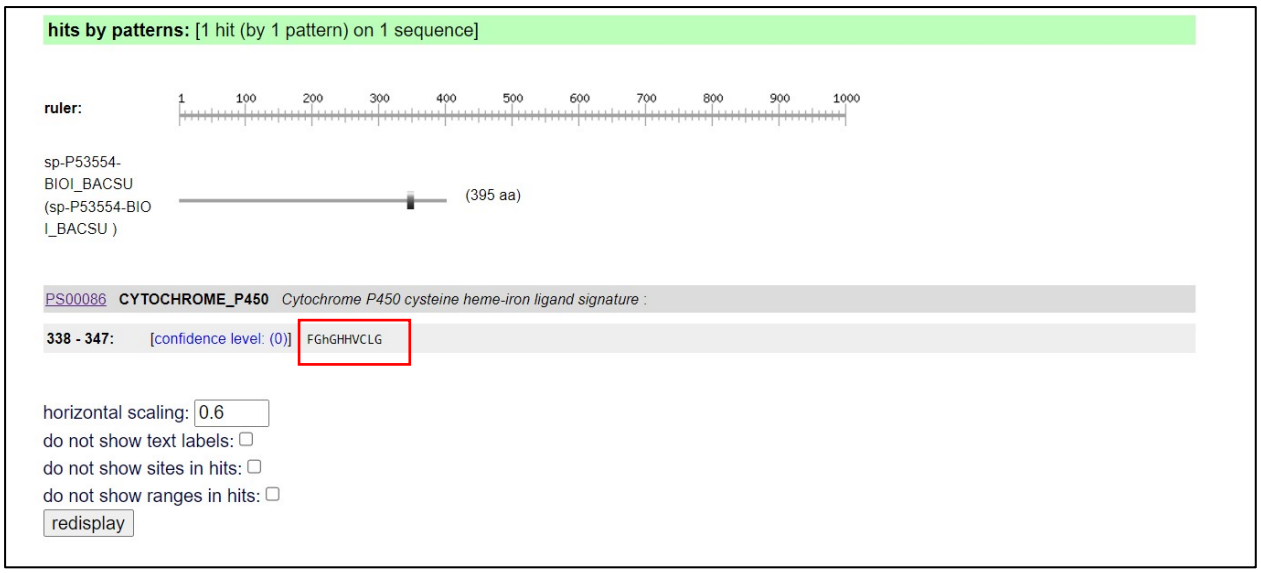


Fig 3b: Results page for the Quick Scan of ScanProSite using the sequence and retrieving the decoded sequence

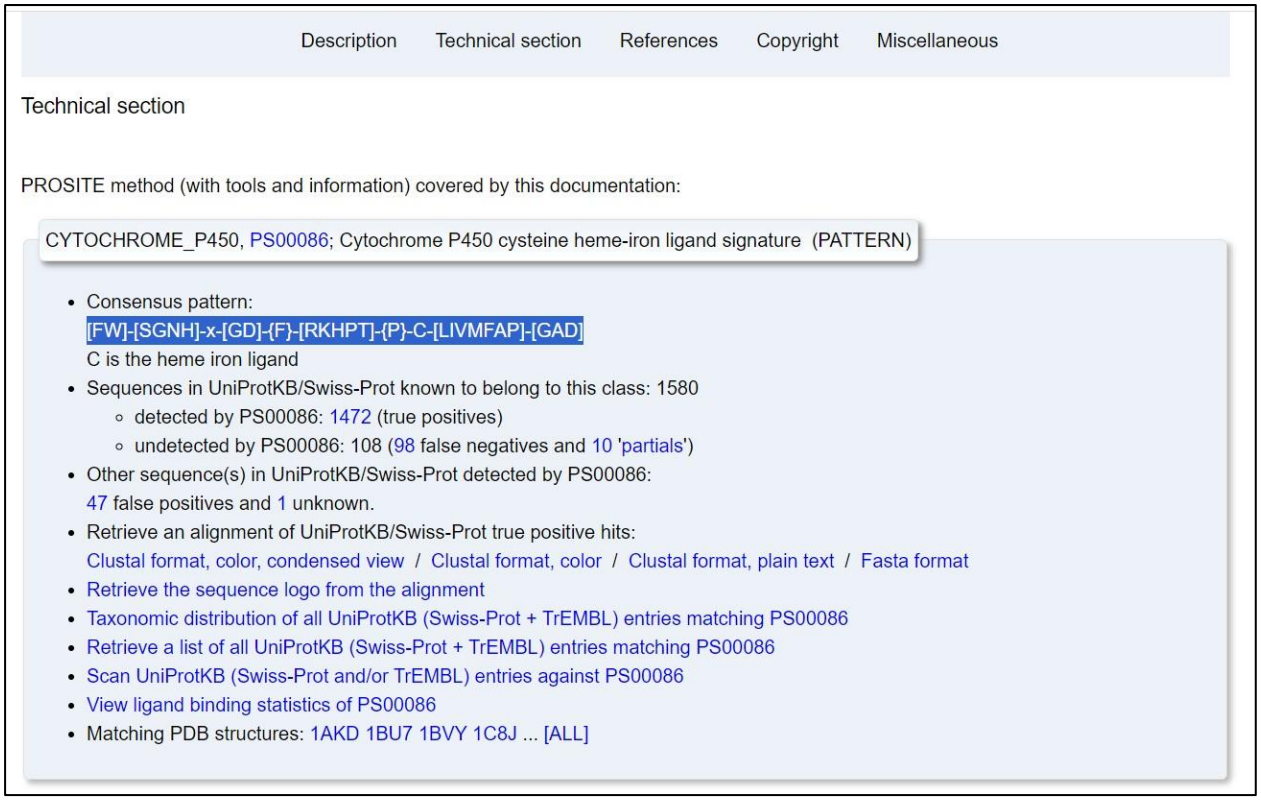


Figure 3c: Consensus pattern for the FASTA sequence

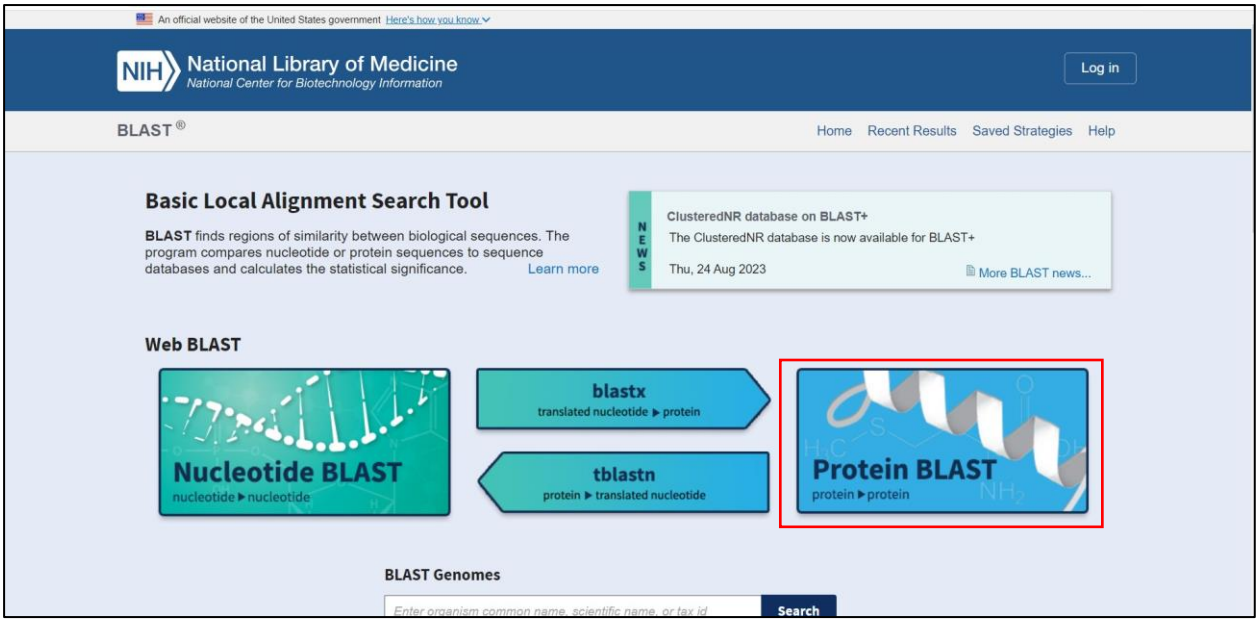


Figure 4: Homepage of Basic Local Alignment Search Tool (BLAST)

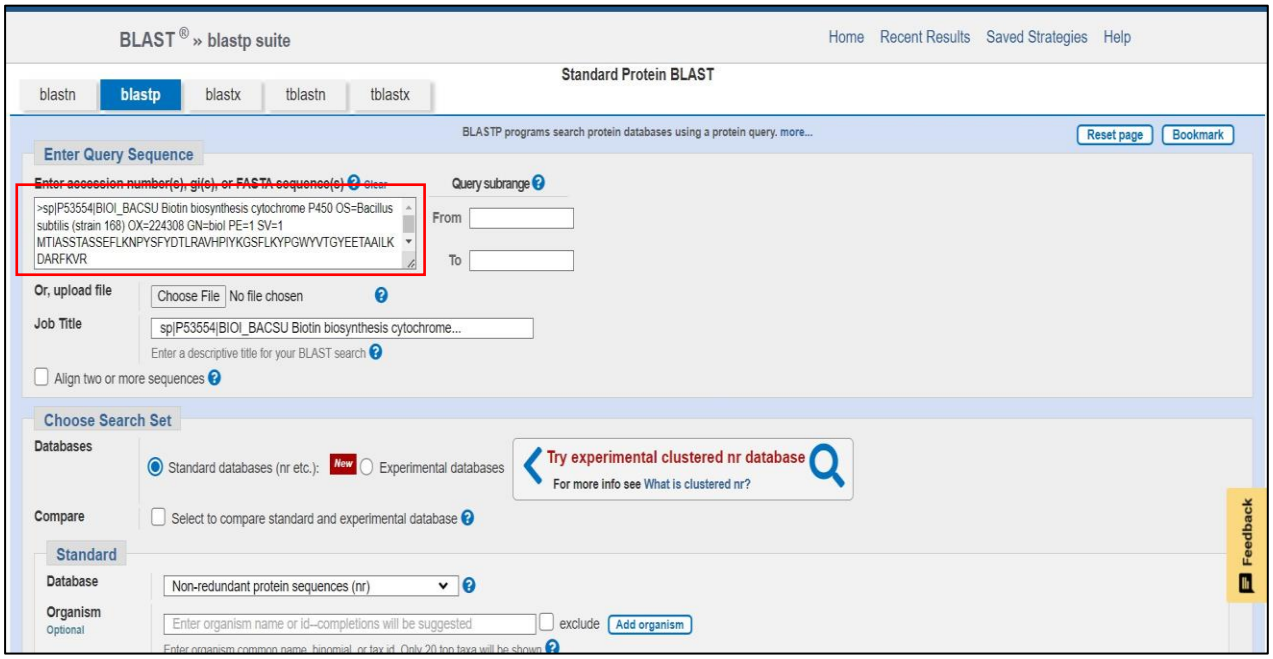


Figure 5: Pasting the FASTA sequence in 'Enter query sequence' box

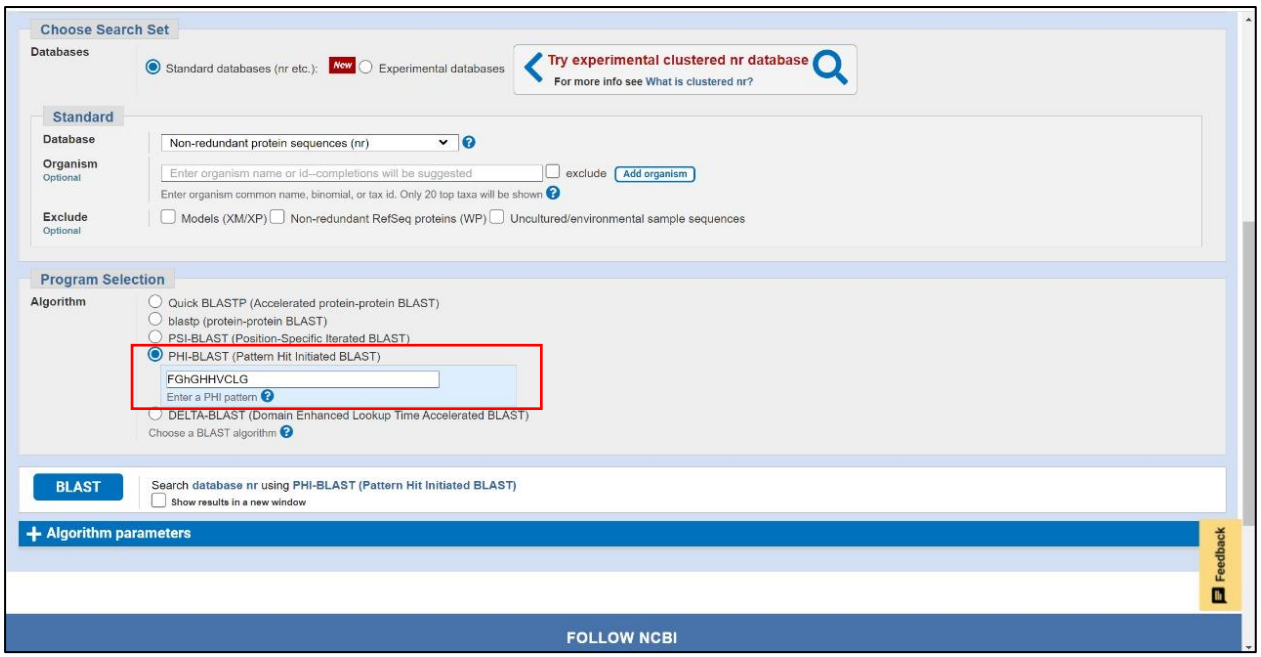


Fig 5a: Paste the decoded pattern from ProSite in 'Enter a PHI pattern' box

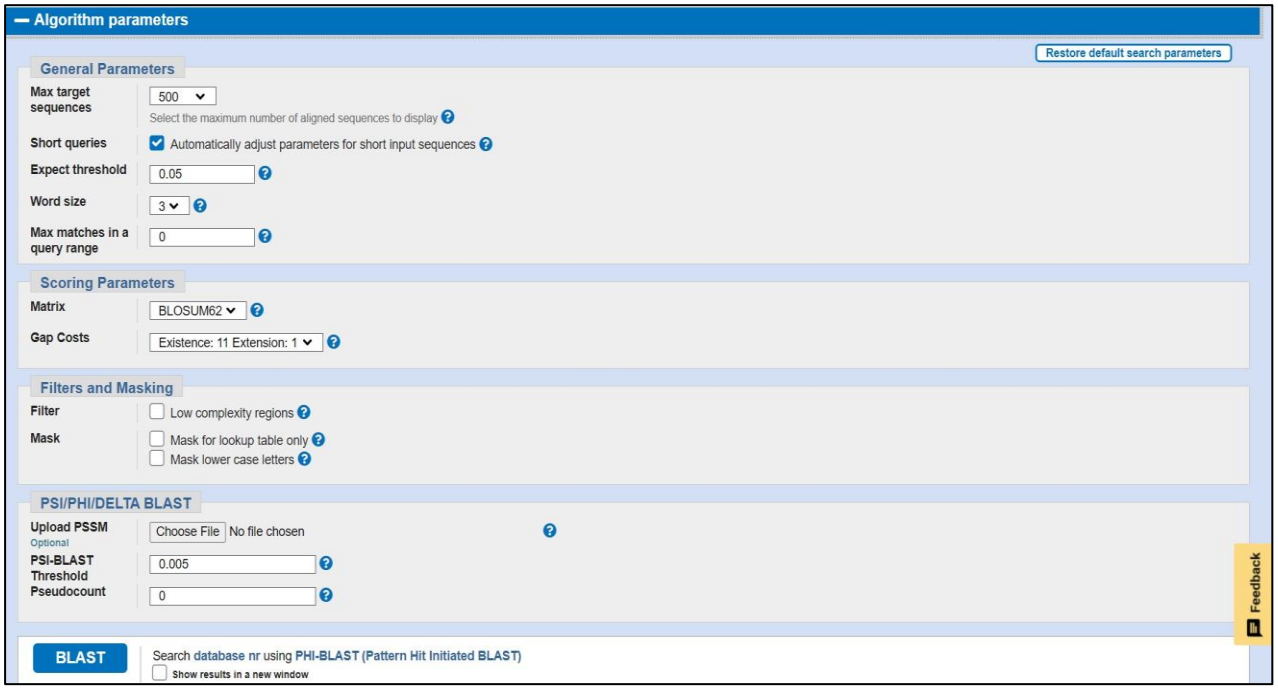


Figure 5b: Setting the parameters for running BLAST Tool

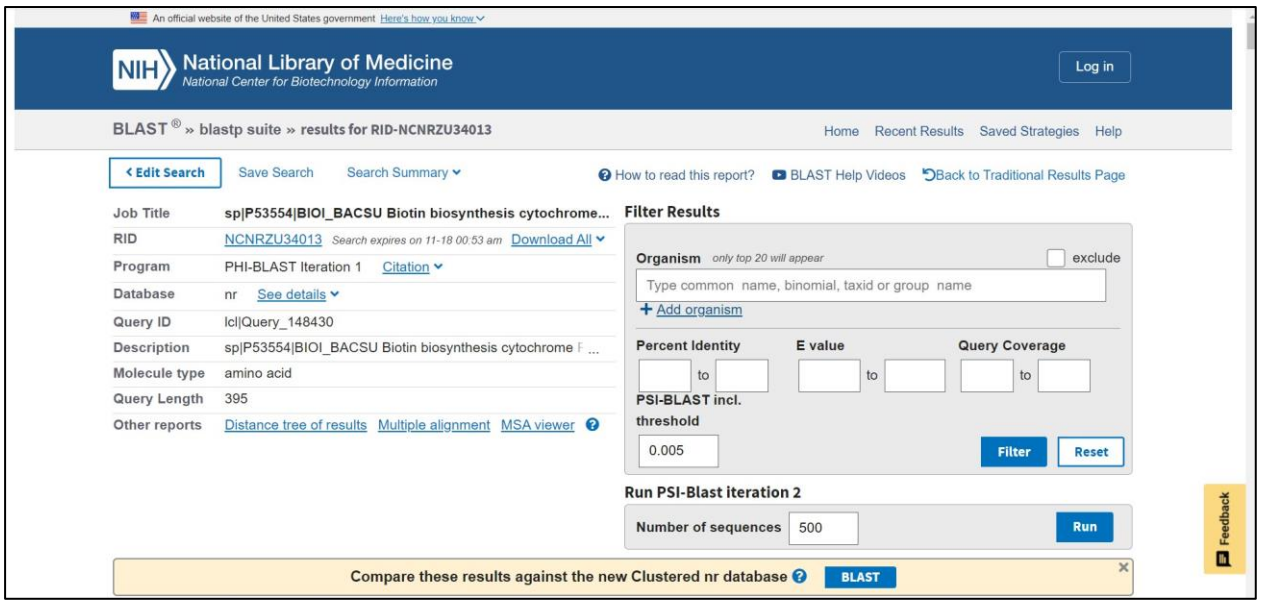
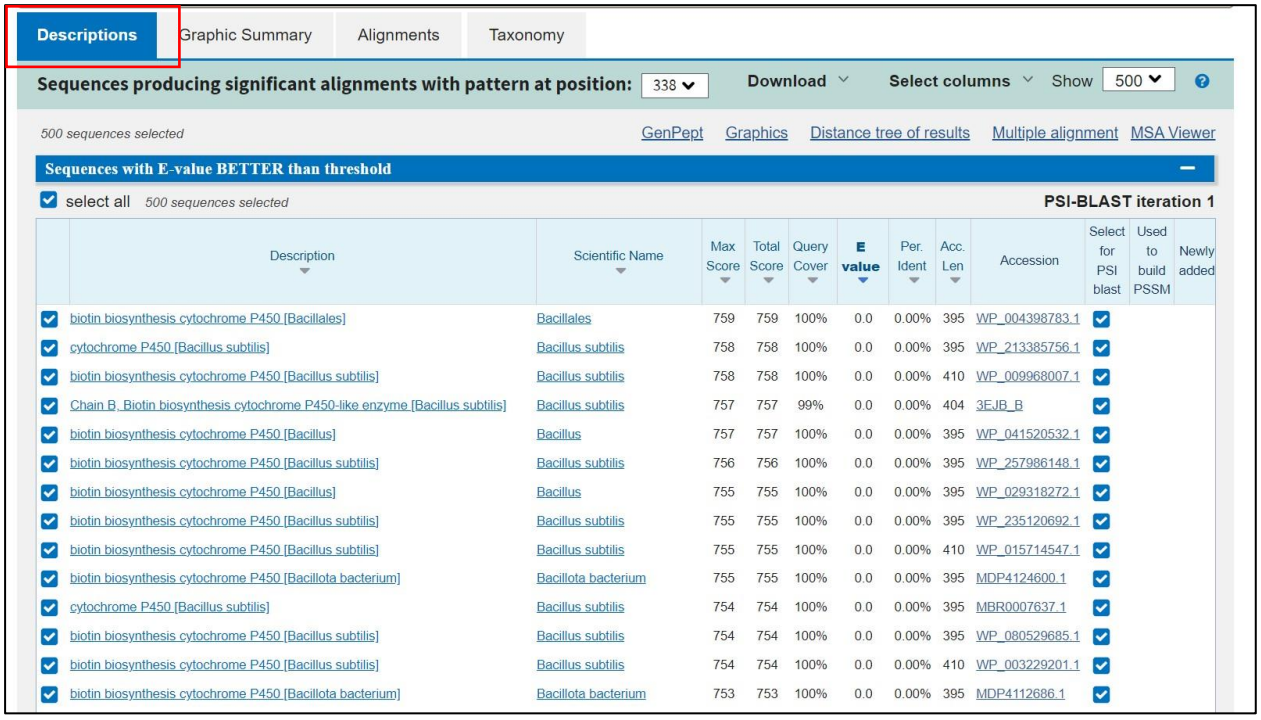


Figure 6: Results obtained after running BLAST tool



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
biotin biosynthesis cytochrome P450 [Bacillales]	Bacillales	759	759	100%	0.0	0.00%	395	WP_004398783.1	✓		
cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	758	758	100%	0.0	0.00%	395	WP_213385756.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	758	758	100%	0.0	0.00%	410	WP_009968007.1	✓		
Chain B, Biotin biosynthesis cytochrome P450-like enzyme [Bacillus subtilis]	Bacillus subtilis	757	757	99%	0.0	0.00%	404	3EJB_B	✓		
biotin biosynthesis cytochrome P450 [Bacillus]	Bacillus	757	757	100%	0.0	0.00%	395	WP_041520532.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	756	756	100%	0.0	0.00%	395	WP_257986148.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus]	Bacillus	755	755	100%	0.0	0.00%	395	WP_029318272.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	755	755	100%	0.0	0.00%	395	WP_235120692.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	755	755	100%	0.0	0.00%	410	WP_015714547.1	✓		
biotin biosynthesis cytochrome P450 [Bacillota bacterium]	Bacillota bacterium	755	755	100%	0.0	0.00%	395	MDP4124600.1	✓		
cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	754	754	100%	0.0	0.00%	395	MBR0007637.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	754	754	100%	0.0	0.00%	395	WP_080529685.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	754	754	100%	0.0	0.00%	410	WP_003229201.1	✓		
biotin biosynthesis cytochrome P450 [Bacillota bacterium]	Bacillota bacterium	753	753	100%	0.0	0.00%	395	MDP4112686.1	✓		

Figure 7: Result for Description section of query

Figure 8: Result for Graphic Summary section

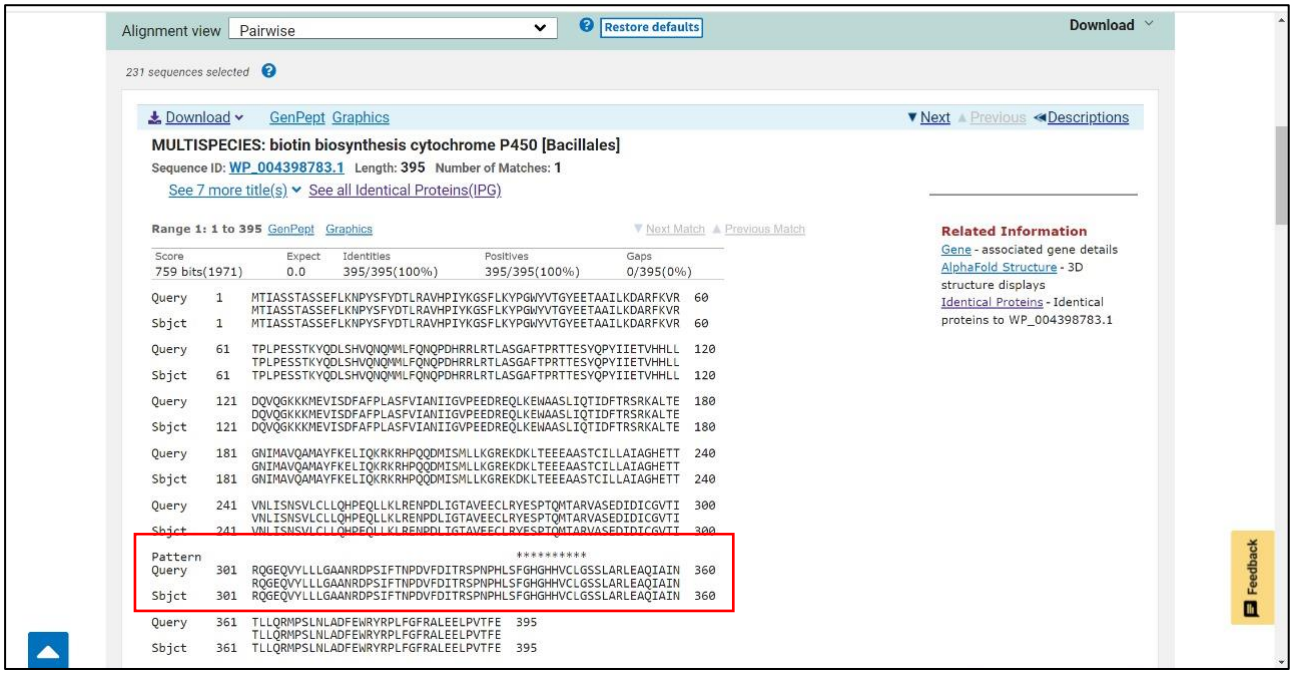


Figure 9: Result for Alignment Section

Descriptions | Graphic Summary | Alignments | Taxonomy

Reports: Lineage | Organism | Taxonomy

100 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
root			334	
synthetic construct	other sequences	1244	13	synthetic construct hits
Homo sapiens	primates	1239	236	Homo sapiens hits
Pongo abelii	primates	1239	5	Pongo abelii hits
Gorilla gorilla gorilla	primates	1229	1	Gorilla gorilla gorilla hits
Pan paniscus	primates	1228	1	Pan paniscus hits
Pan troglodytes	primates	1228	3	Pan troglodytes hits
Pongo pygmaeus	primates	1219	1	Pongo pygmaeus hits
Nomascus leucogenys	primates	1211	1	Nomascus leucogenys hits
Hylobates moloch	primates	1211	1	Hylobates moloch hits
Symphalangus syndactylus	primates	1206	1	Symphalangus syndactylus hits
unidentified	unclassified sequences	1188	2	unidentified hits
Macaca mulatta	primates	1175	4	Macaca mulatta hits
Macaca fascicularis	primates	1175	5	Macaca fascicularis hits
Macaca thibetana thibetana	primates	1174	1	Macaca thibetana thibetana hits
Theropithecus gelada	primates	1173	1	Theropithecus gelada hits
Macaca nemestrina	primates	1172	1	Macaca nemestrina hits

Feedback

Figure 10: Result for Taxonomy section based on 'Lineage'

Descriptions | Graphic Summary | Alignments | Taxonomy

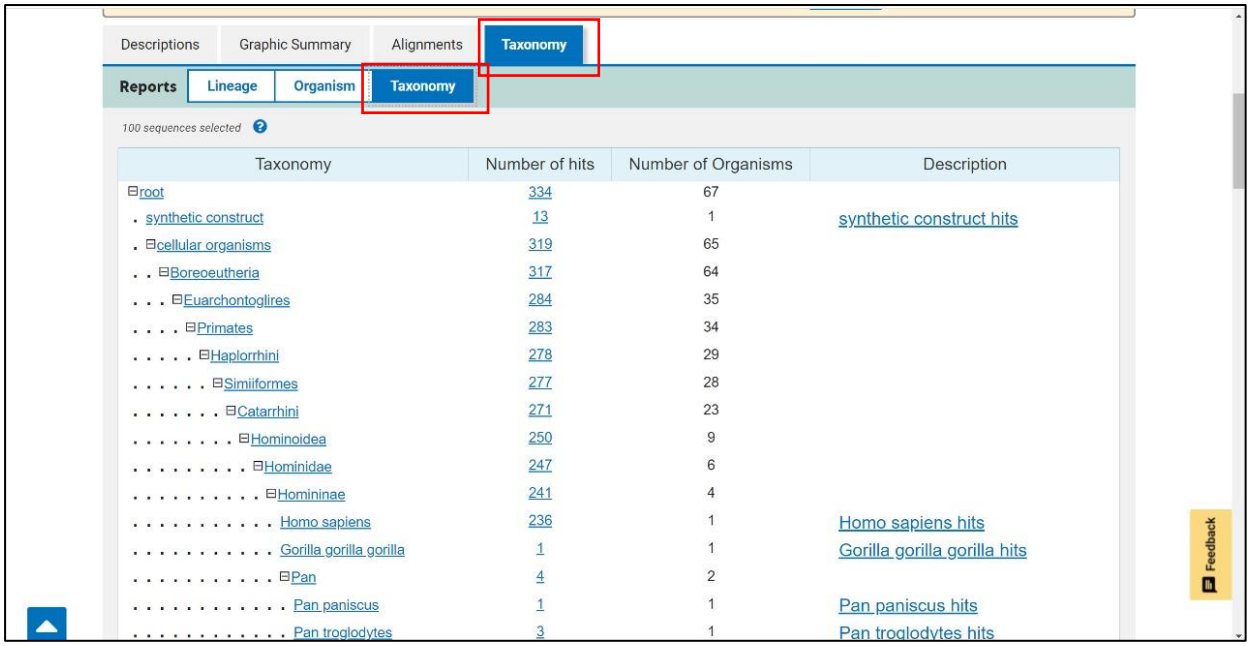
Reports: Lineage | Organism | Taxonomy

100 sequences selected

Description	Score	E value	Accession
synthetic construct [other sequences]	▼Next	▲Previous	◄◄First
serum albumin-interferon alpha 1 fusion protein, partial [synthetic construct]	1244	0.0	AGI02589
albumin, partial [synthetic construct]	1239	0.0	AAX36126
albumin [synthetic construct]	1239	0.0	ABM82340
serum albumin [synthetic construct]	1220	0.0	AIC32938
HSA-cIFN [synthetic construct]	1195	0.0	QCO95453
HSA-GGGGS-GH fusion protein, partial [synthetic construct]	1192	0.0	AFO84000
IL-1Ra-GGGGS-HSA fusion protein, partial [synthetic construct]	1191	0.0	AEL88488
HSA-GGGGS-IL-1Ra fusion protein, partial [synthetic construct]	1191	0.0	AEZ51871
human serum albumin and interferon-alpha2b fusion protein, partial [synthetic construct]	1190	0.0	QNI40628
HSA-GGGGS-PTH(1-34), partial [synthetic construct]	1189	0.0	AER13700
serum albumin, partial [synthetic construct]	1188	0.0	AIC32937
somatostatin (SST) doublet/albumin fusion protein [synthetic construct]	1186	0.0	UTT97830
human serum albumin mutein, partial [synthetic construct]	1185	0.0	QNI40627
Homo sapiens (human) [primates]	▼Next	▲Previous	◄◄First
albumin preproprotein [Homo sapiens]	1239	0.0	NP_000468
RecName: Full=Albumin; Flags: Precursor [Homo sapiens]	1239	0.0	P02768
Chain A, SERUM ALBUMIN [Homo sapiens]	1239	0.0	4BKE_A

Feedback

Figure 11: Result for Taxonomy section based on 'Organism'



Descriptions | Graphic Summary | Alignments | Taxonomy

Reports | Lineage | Organism | Taxonomy

100 sequences selected

Taxonomy	Number of hits	Number of Organisms	Description
⊟root	334	67	
. synthetic construct	13	1	synthetic construct hits
. ⊟cellular organisms	319	65	
. . ⊟Boreoeutheria	317	64	
. . . ⊟Euarchontoglires	284	35	
. . . . ⊟Primates	283	34	
. . . . . ⊟Haplorrhini	278	29	
. . . . . ⊟Simiiformes	277	28	
. . . . . ⊟Catarrhini	271	23	
. . . . . ⊟Hominoidea	250	9	
. . . . . ⊟Hominidae	247	6	
. . . . . ⊟Homininae	241	4	
. . . . . Homo sapiens	236	1	Homo sapiens hits
. . . . . Gorilla gorilla gorilla	1	1	Gorilla gorilla gorilla hits
. . . . . ⊟Pan	4	2	
. . . . . Pan paniscus	1	1	Pan paniscus hits
. . . . . Pan troglodytes	3	1	Pan troglodytes hits

Feedback

Figure 12: Result for Taxonomy section based on 'Taxonomy'

RESULTS:

Pattern-Hit Initiated BLAST (PHI-BLAST) tool is a variant of the Basic Local Alignment Search Tool (BLAST) algorithm, specifically designed for detecting distant relationships between protein sequences and identifying domains of potential functional significance within sequences. The tool was used to studied query where it is able to detect the pattern in the organisms which confirms the identification of remote homologs or conserved domains for the query protein sequences.

CONCLUSION:

PHI-BLAST is widely used in bioinformatics, particularly for analyzing protein sequences to identify conserved domains, motifs, or functional signatures. It aids in understanding evolutionary relationships between proteins and assists in annotating sequences with functional information based on conserved patterns. Its ability to focus the alignment and construction of the PSSM around a motif provides a valuable approach for researchers and bioinformaticians working in the field of protein analysis.

REFERENCES:

1. ResearchGate. (2023). BLAST Algorithm. https://www.reseatchgate.net/publication/230503487
2. Zheng Zhang, Webb Miller, Alejandro A. Schäffer, Thomas L. Madden, David J. Lipman, Eugene V. Koonin, Stephen F. Altschul, Protein sequence similarity searches using patterns as seeds, Nucleic Acids Research, Volume 26, Issue 17, 1 September 1998, Pages 3986–3990, https://doi.org/10.1093/nar/26.17.3986
3. Sancho J. Flavodoxins: sequence, folding, binding, function and beyond. Cell Mol Life Sci. 2006 Apr;63(7-8):855-64. doi: 10.1007/s00018-005-5514-4. PMID: 16465441. https://pubmed.ncbi.nlm.nih.gov/16465441

DATE: 01/11/23

WEBLEM 6(E)
EMBOSS NEEDLE – GLOBAL PAIRWISE SEQUENCE ALIGNMENT
(URL: https://www.ebi.ac.uk/Tools/psa/emboss_needle/)

AIM:

To explore and compare the protein sequences of 'Myosin' from two organisms *Gallus gallus* (UniProt ID: Q90623) and *Mus musculus* (UniProt ID: F8VQB6) by performing global pairwise sequence alignment using EMBOSS Needle Tool.

INTRODUCTION:

The European Molecular Biology Open Software Suite, or EMBOSS, is a part of the European Bioinformatics Institute (EBI). One of the prominent tools of EMBOSS is EMBOSS Needle, which is based on the Needleman-Wunsch algorithm. The Needleman-Wunsch algorithm was developed by Saul B. Needleman and Christian D. Wunsch in 1970 for global sequence alignment. It works on the principle of dividing the large problem into a series of smaller problems and uses the solutions to the smaller problems to find an optimal solution to the larger problem, assigning a score to every possible alignment and finding all possible alignments having the highest score.

The unique feature of the EMBOSS Needle tool is that it finds the alignment with the maximum possible score where the score of an alignment is equal to the sum of the matches taken from the scoring matrix, minus penalties arising from opening and extending gaps in the aligned sequences. The substitution matrix and gap opening and extension penalties are user-specified. A penalty is subtracted from the score for each gap opened (Gap insertion penalty) and a penalty is subtracted from the score for the extension of the inserted gaps (Gap extension penalty). Typically, the cost of extending a gap is set to be 5-10 times lower than the cost for opening a gap.

Penalty for a gap of n positions is calculated using the following formula:

Gap at n^{th} position = gap opening penalty + (n - 1) * gap extension penalty

Myosin:

Myosin is a motor protein with a primary role in muscle contraction, interacting with actin filaments to generate force and movement. Beyond muscles, myosin participates in cell motility, cell division, intracellular transport, and maintenance of cell shape, making it a crucial component in various cellular processes. The need to analyze myosin with the EMBOSS Needle tool arises from the diverse functions of myosin, which contribute to the dynamic behavior and structural integrity of cells. By analyzing the sequence and structure of myosin, researchers can gain insights into its mechanisms and interactions, which can help develop a deeper understanding of its role in various cellular processes and potentially lead to new therapeutic strategies for muscle and non-muscle related disorders.

METHODOLOGY:

1. Open the UniProt database and search for the query of 'Myosin'.
2. From the results page, open the proteins of interest. Here, *Gallus gallus* (UniProt ID: Q90623) and *Mus musculus* (UniProt ID: F8VQB6).
3. Download the myosin protein sequences of both the organisms in FASTA file format.
4. Open the homepage of EMBOSS Needle tool and paste the sequences in the query box and set the desired parameters. Select the 'SUBMIT' to submit the query.
5. The results page of EMBOSS Needle tool displays the Alignment, Submission Details and View Alignment File. Interpret the results.

OBSERVATIONS:

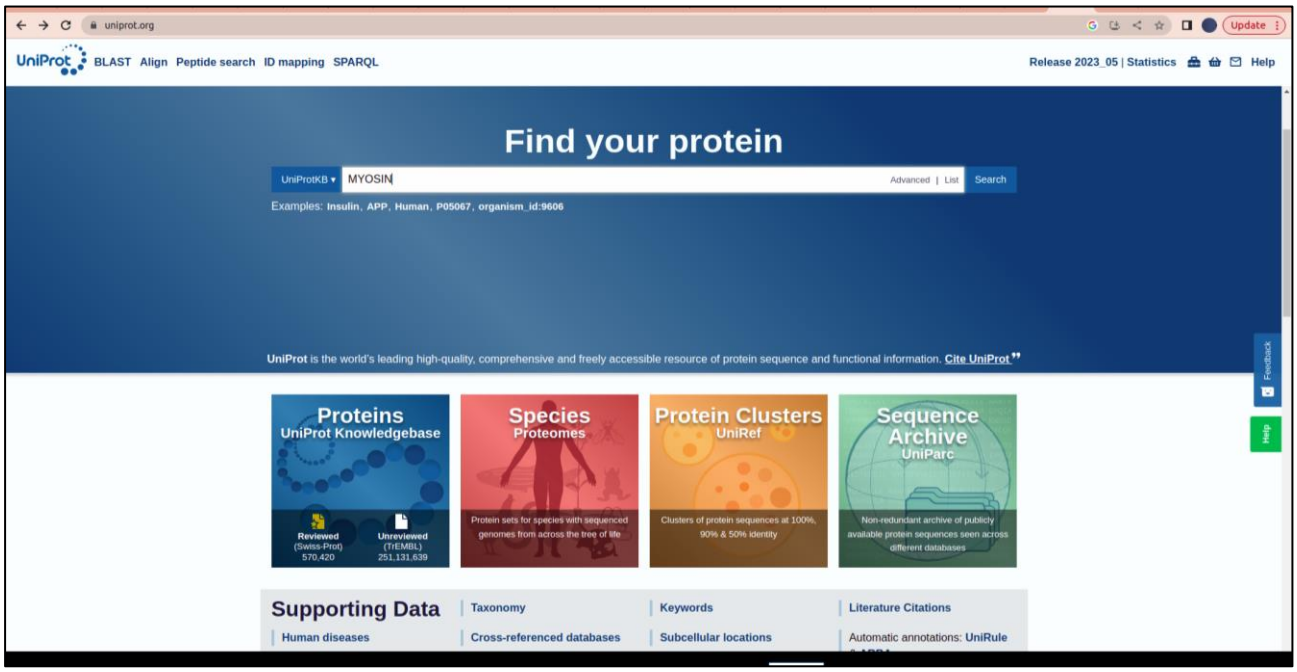


Figure 1: Homepage of the UniProt Database

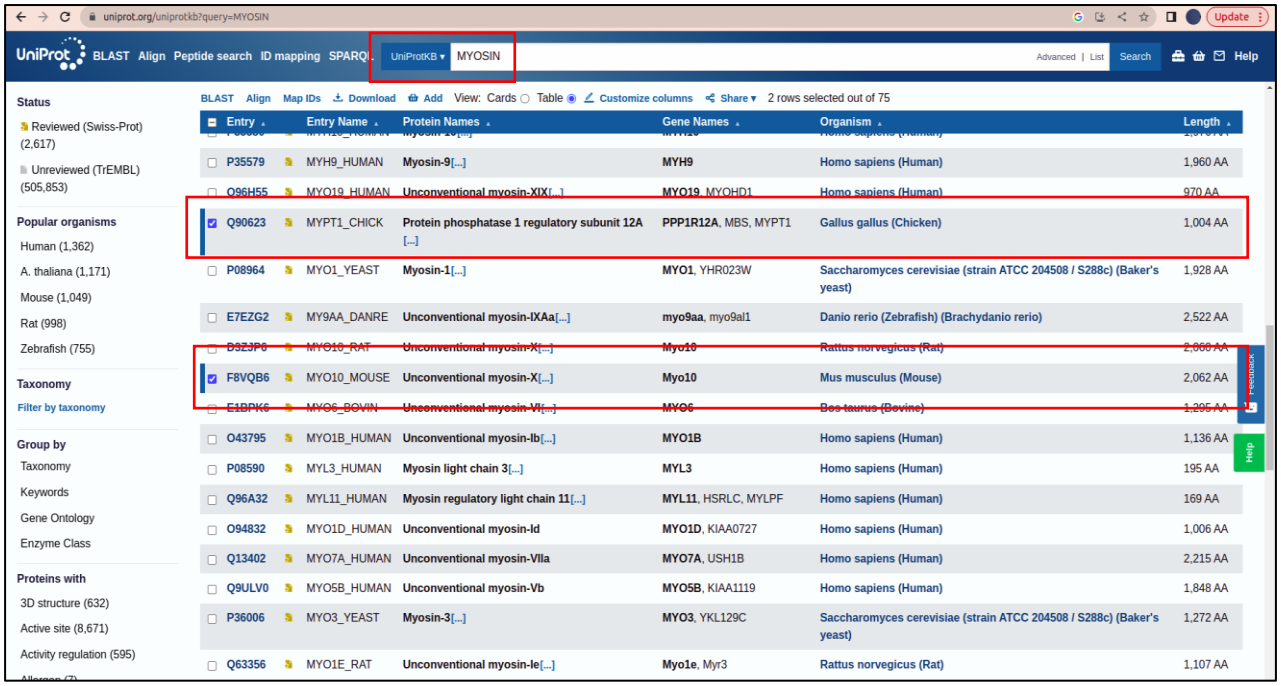


Figure 2: Results page of the UniProt Database for the query of Myosin with selected entries

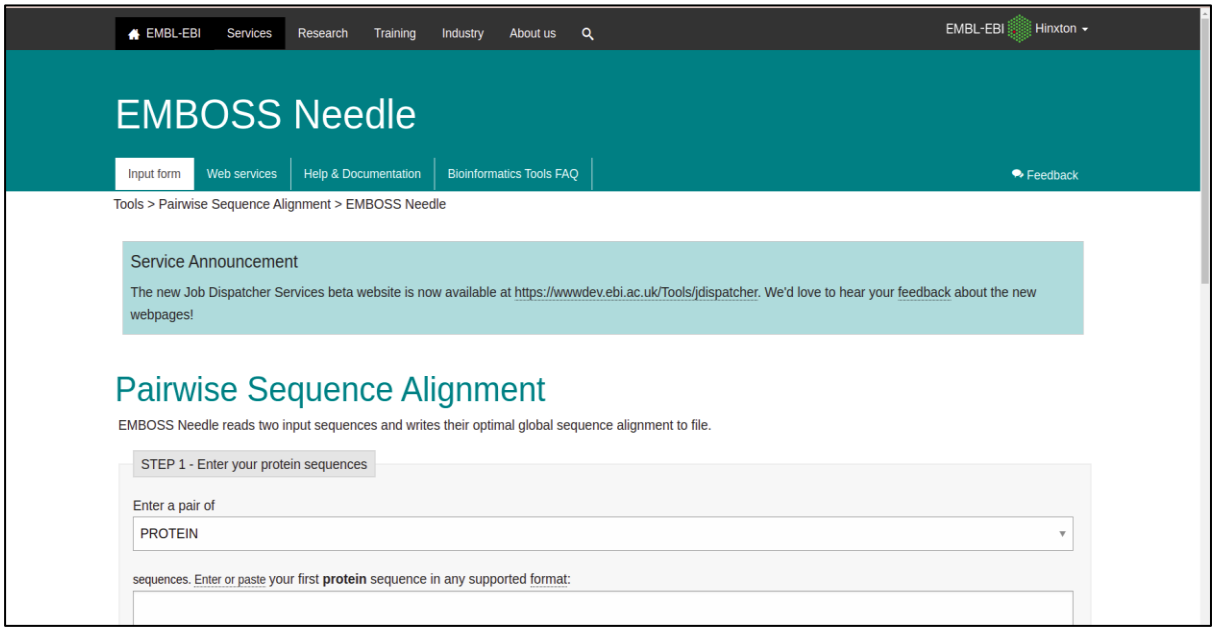


Figure 3: Homepage of EMBOSS Needle Tool

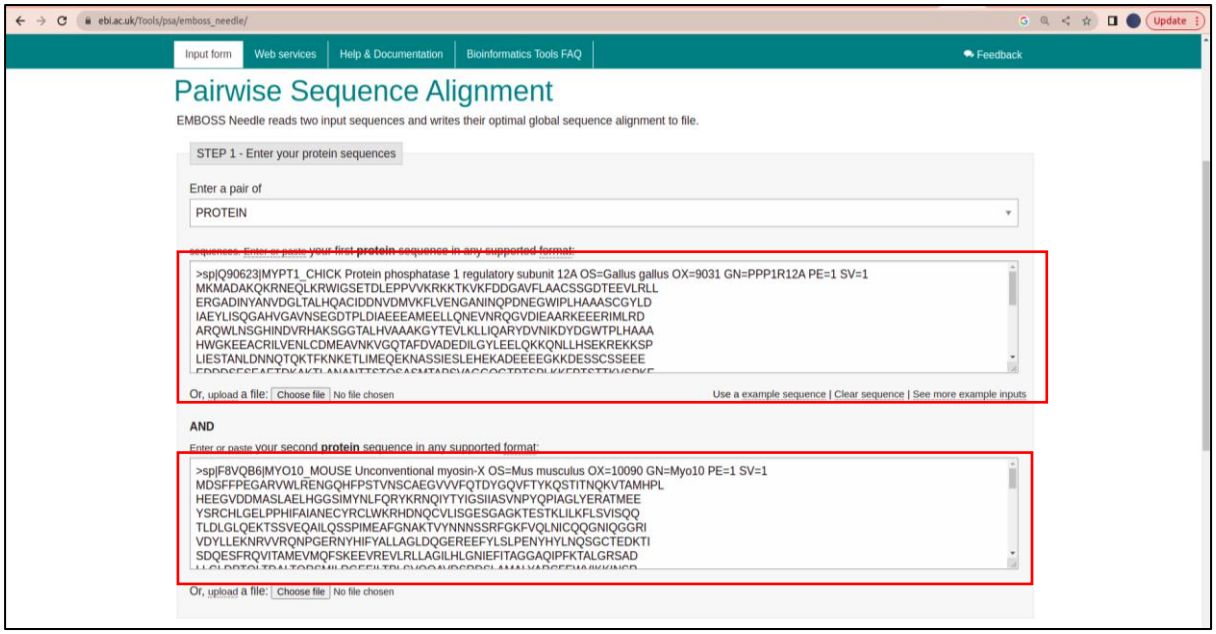


Figure 4: Submission of the protein sequences retrieved from the UniProt Database in the EMBOSS Needle Tool

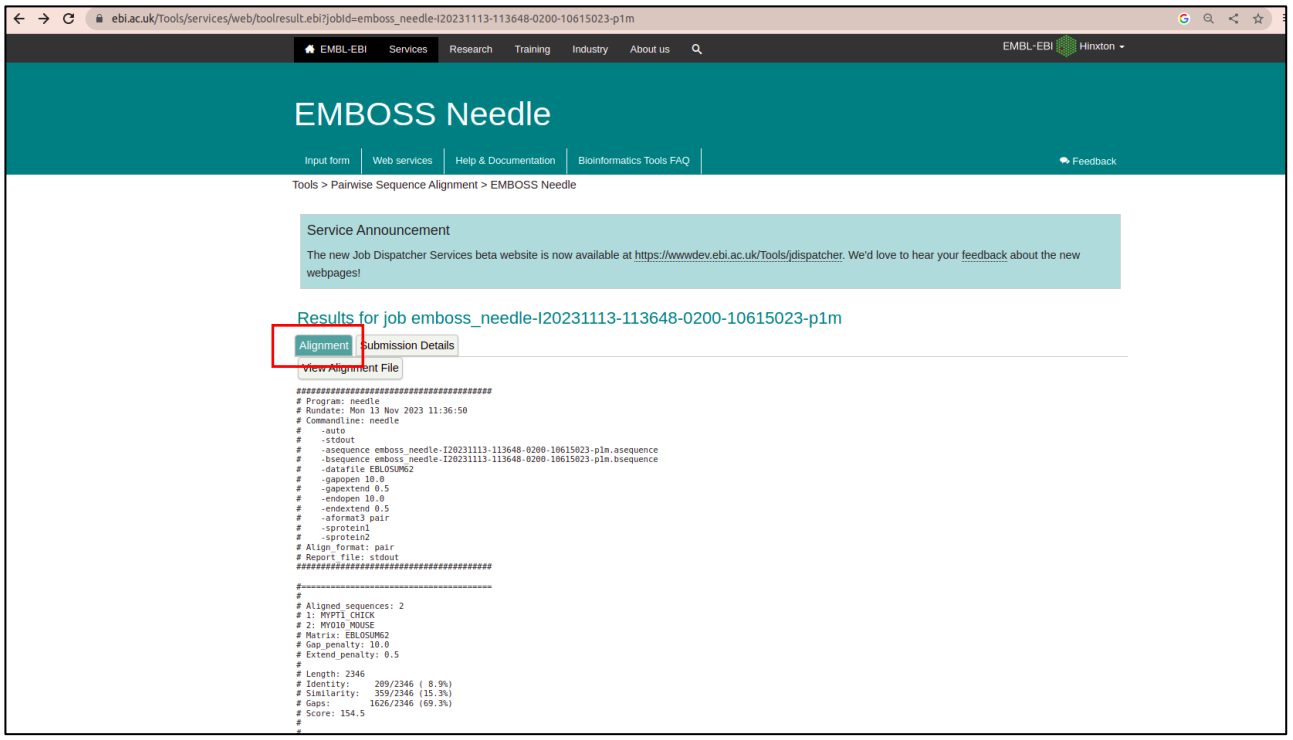


Figure 5: Results page of the submitted query with Alignment option

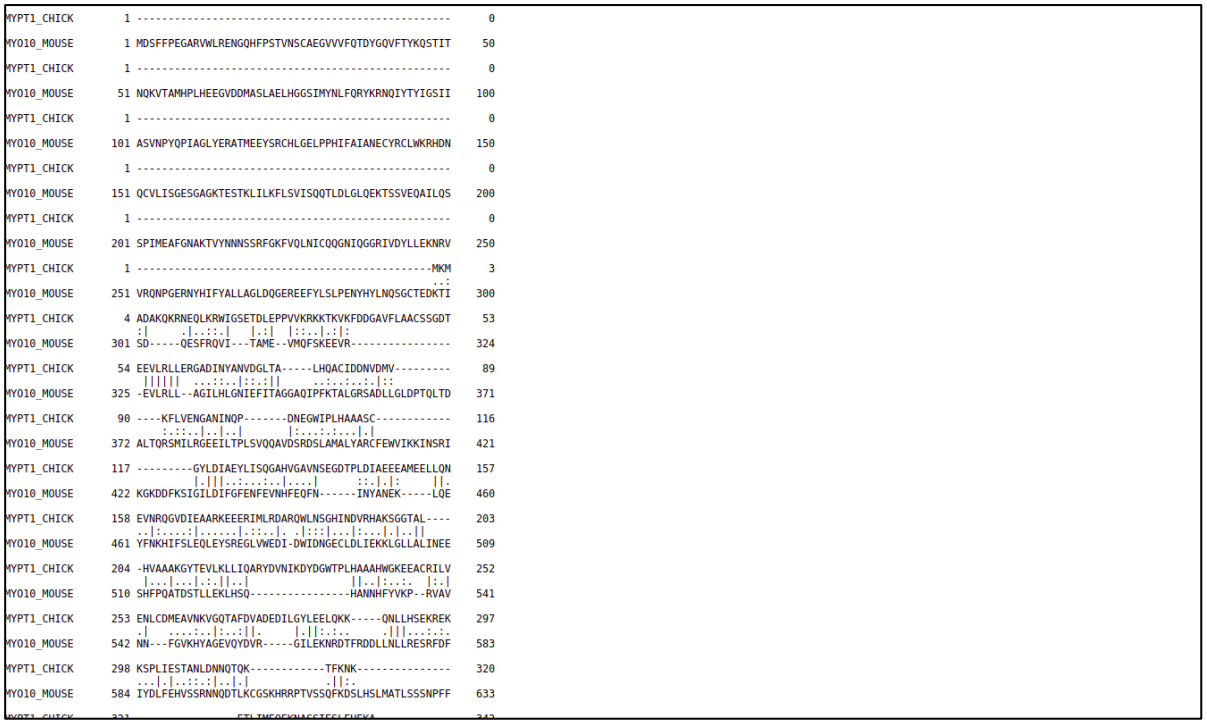


Figure 5a: Results page of the submitted query with Alignment option

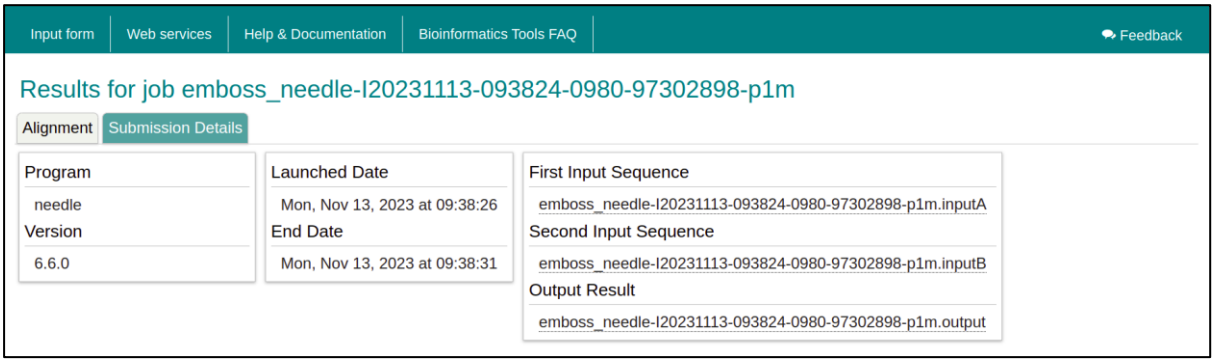


Figure 6: View of submission details

RESULTS:

By exploring global pairwise sequence alignment using the EMBOSS Needle tool, the results were observed and studied for the protein query 'Myosin' in organisms *Gallus gallus* (UniProt ID: Q90623) and *Mus musculus* (UniProt ID: F8VQB6). It was found that in the pairwise alignment of the two organisms, they were not identical upon comparison, as the sequence identity is only 8.9%.

Length	2346
Identity	209/2346 (8.9%)
Similarity	359/2346 (15.3%)
Gaps	1626/2346 (69.3%)
Score	154.5

CONCLUSION:

EMBOSS Needle tool, for Global Pairwise Sequence Alignment, was explored by comparative study of protein 'Myosin' of two different organisms, namely, *Gallus gallus* (UniProt ID: Q90623) and *Mus musculus* (UniProt ID: F8VQB6).

REFERENCES:

1. Needleman, S. B. and Wunsch, C. D. (1970) *J. Mol. Biol*. 48, 443-453. https://www.bioinformatics.nl/cgi-bin/emboss/help/needle
2. Robert S. Adelstein, James R. Sellers, in *Biochemistry of Smooth Muscle Contraction*, 1996. https://doi.org/10.1016/B978-0-12-801387-8.00003-X

DATE: 01/11/23

WEBLEM 6(F)
EMBOSS WATER – LOCAL PAIRWISE SEQUENCE ALIGNMENT
(URL: https://www.ebi.ac.uk/Tools/psa/emboss_water/)

AIM:

To explore and compare the protein sequences of 'collagen' in two organisms, *Rattus norvegicus* (UniProt ID: P05539) and *Homo sapiens* (UniProt ID: P08572), by performing local pairwise sequence alignment using the EMBOSS Water tool.

INTRODUCTION:

The European Molecular Biology Open Software Suite, or EMBOSS, is a part of the European Bioinformatics Institute (EBI). One of the prominent tools of EMBOSS is EMBOSS Water, which is based on the Smith-Waterman algorithm. Smith-Waterman algorithm was developed by Temple F. Smith and Michael S. Waterman in 1981 and is used for local sequence alignment, which finds the best subsequence match between two sequences by comparing all possible pairs of subsequences. The unique aspect of the EMBOSS Water tool is that it uses a speed-accelerated version of the Smith-Waterman method to determine the local alignment of a sequence with one or more other sequences. By examining every potential alignment and choosing the best one, dynamic programming techniques guarantee the best possible local alignment. To do this, a scoring matrix with values for each potential residue or nucleotide match is incorporated.

The EMBOSS Water tool employs a modified Smith-Waterman algorithm with speed enhancements to compute the local alignment of one or more sequences. Users have the flexibility to specify the gap insertion penalty, gap extension penalty, and substitution matrix for calculating alignments. The output is a standard EMBOSS alignment file. Identity refers to the percentage of identical matches between two sequences over the entire reported aligned region, inclusive of any length gaps. Similarly, similarity represents the percentage of matches between the two sequences over the length of the reported aligned region, considering any gaps.

Collagen:

The most prevalent protein in the body, collagen, is found in various connective tissues such as the skin, tendons, bones, and ligaments. Its inherent stiffness and resistance to stretching contribute significantly to providing structural support within the extracellular space of connective tissues. Understanding collagen's structure, function, and its implications in various diseases and conditions, including autoimmune disorders like rheumatoid arthritis, lupus, dermatomyositis, and scleroderma, is crucial. These conditions can adversely affect collagen, highlighting the importance of in-depth research.

The EMBOSS Water tool serves as a valuable resource in this pursuit. It is a pairwise sequence alignment program designed to determine the local alignment of one or more sequences. The tool utilizes a modified version of the Smith-Waterman technique, offering faster results for

researchers. By employing the EMBOSS Water tool to analyze collagen, researchers can gain deeper insights into its molecular makeup and its role in health and disease.

METHODOLOGY:

1. Open the UniProt database and search for the query of 'Collagen'.
2. From the results page, open the proteins of interest. Here, *Rattus norvegicus* (UniProt ID: P05539) and *Homo sapiens* (UniProt ID: P08572).
3. Download the collagen protein sequences of both the organisms in FASTA canonical file format.
4. Open the homepage of EMBOSS Water tool and paste the sequences in the query box and set the desired parameters. Select the 'SUBMIT' to submit the query.
5. The results page of EMBOSS Water tool displays the Alignment, Submission Details and View Alignment File. Interpret the results.

OBSERVATIONS:

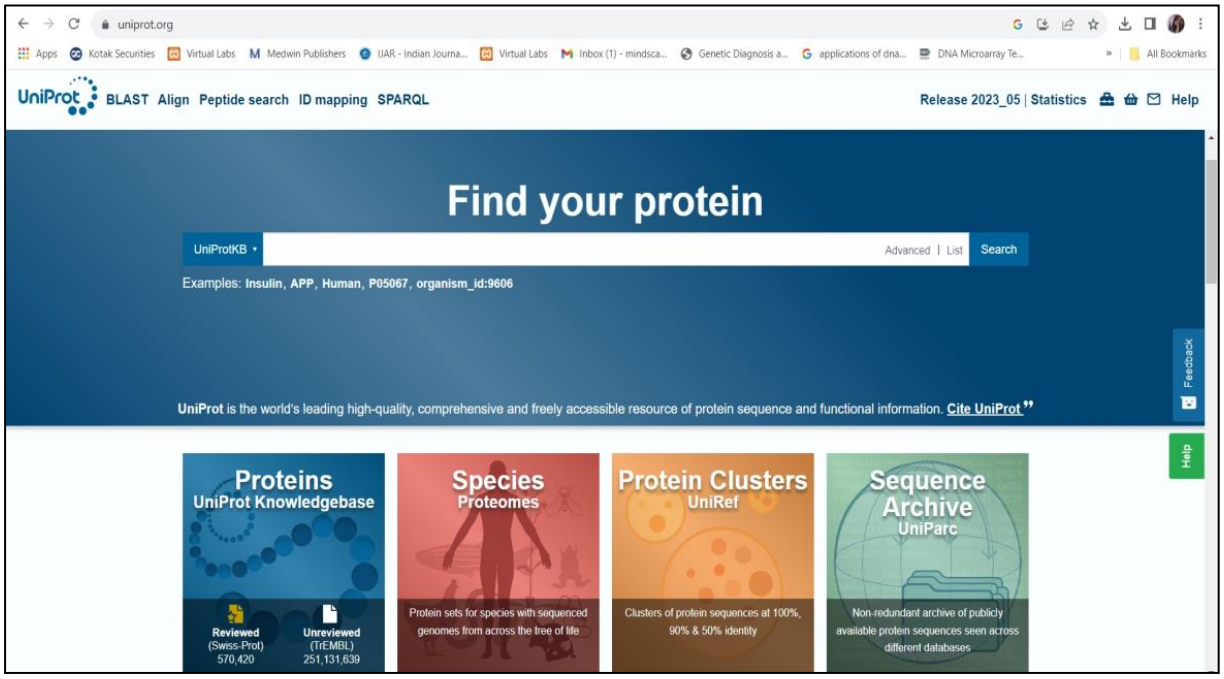


Figure 1: Homepage of the UniProt database

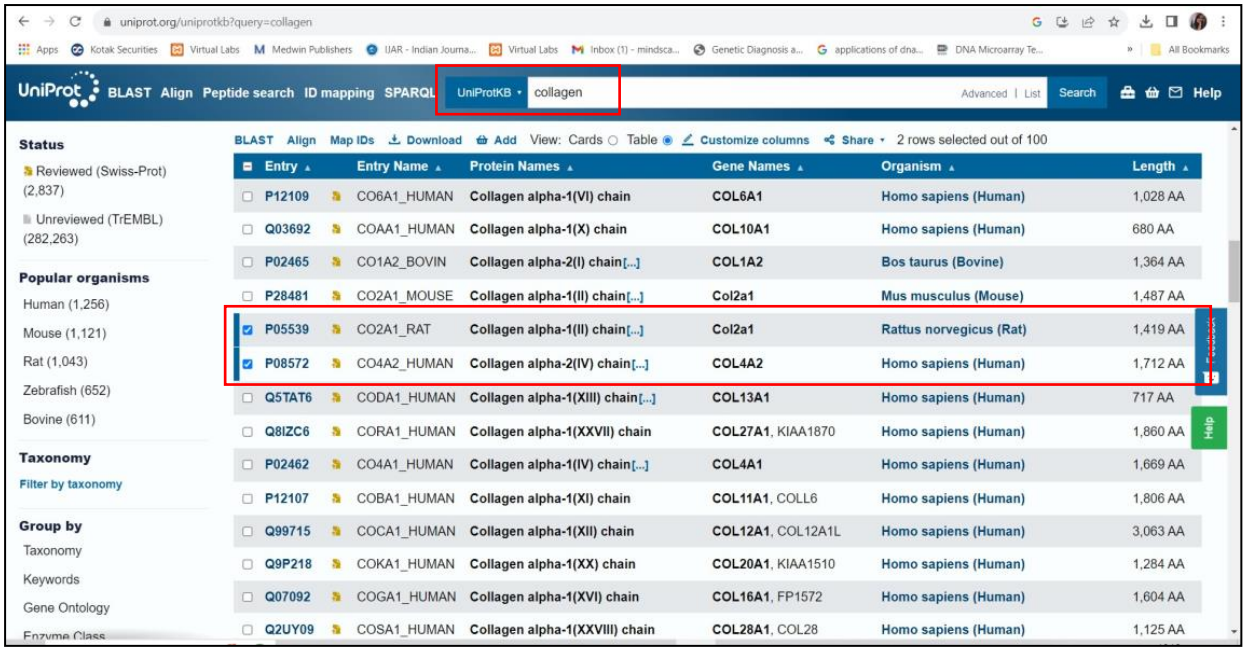


Figure 2: Results page of the UniProt Database for the query of collagen with selected entries

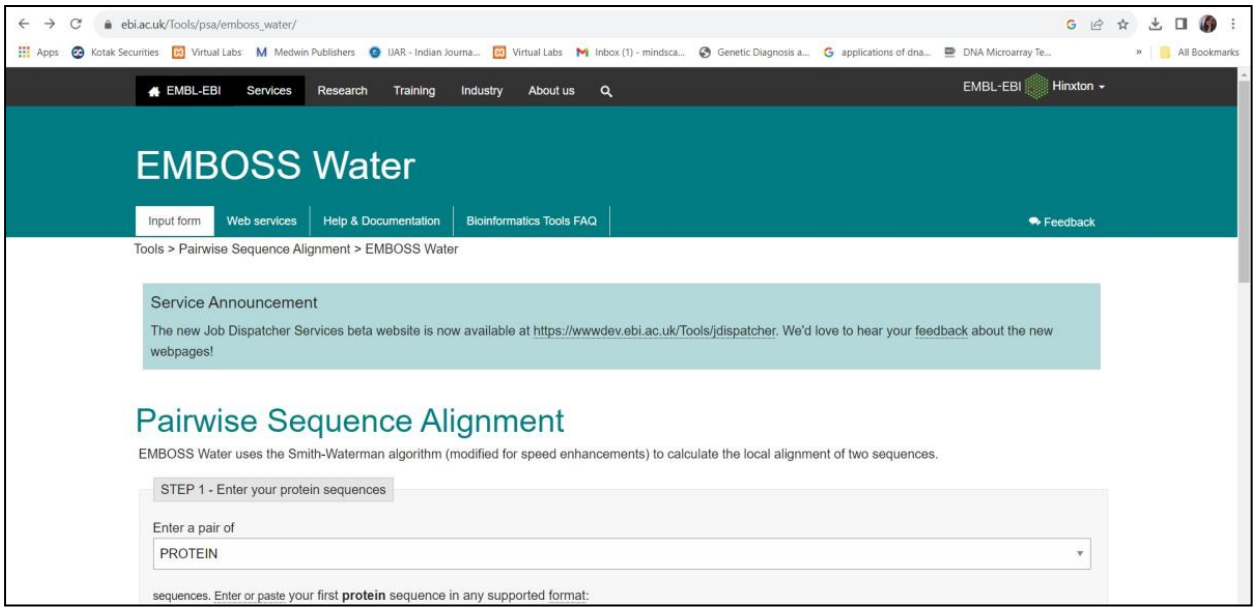


Figure 3: Homepage of EMBOSS Water Tool

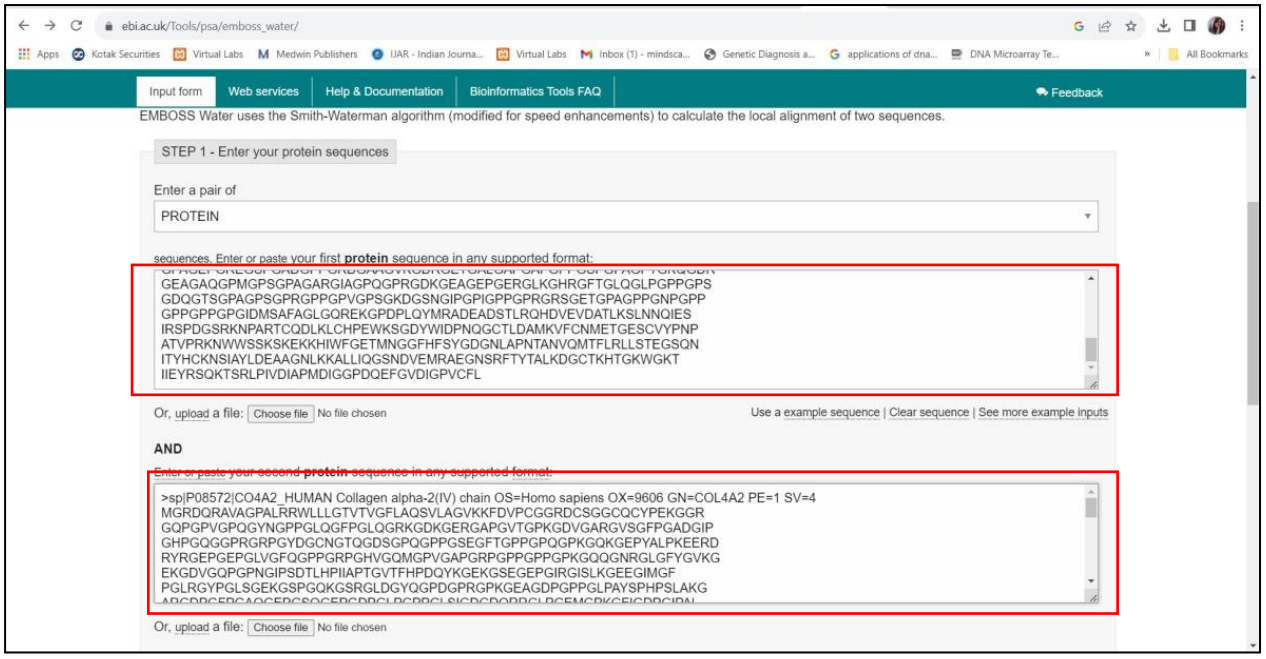


Figure 4: Submission of the protein sequences retrieved from the UniProt Database in the EMBOSS Water Tool

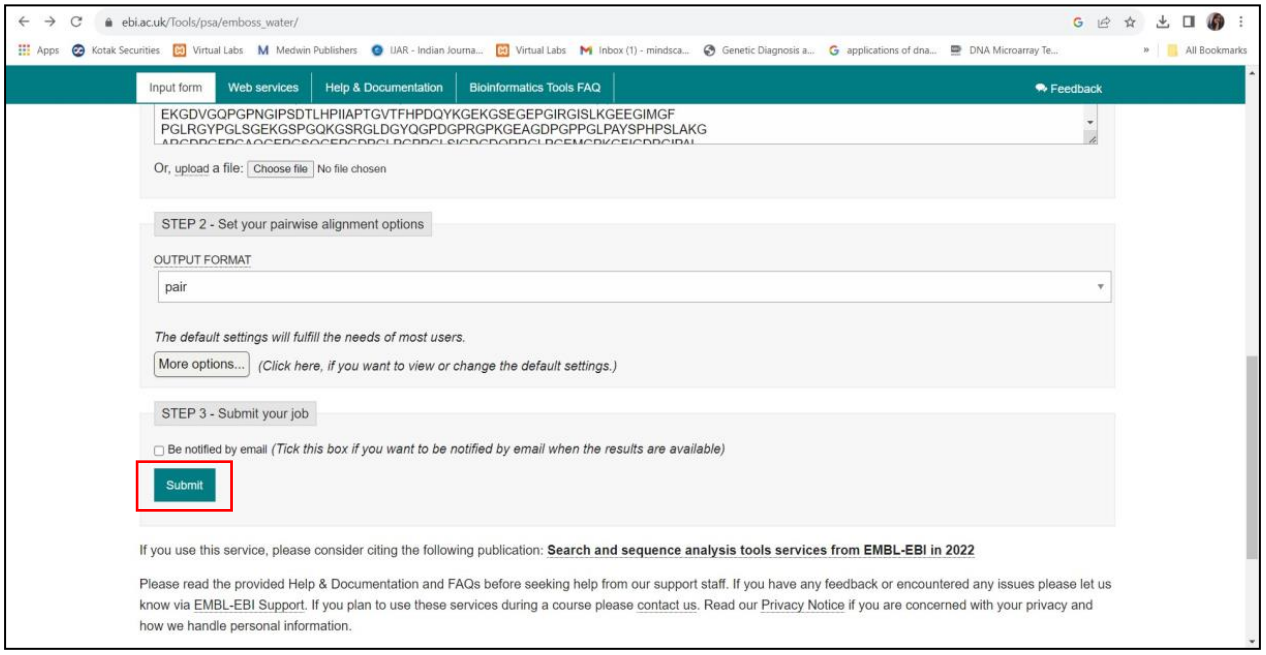


Figure 5: Submission of the query to the EMBOSS Water Tool

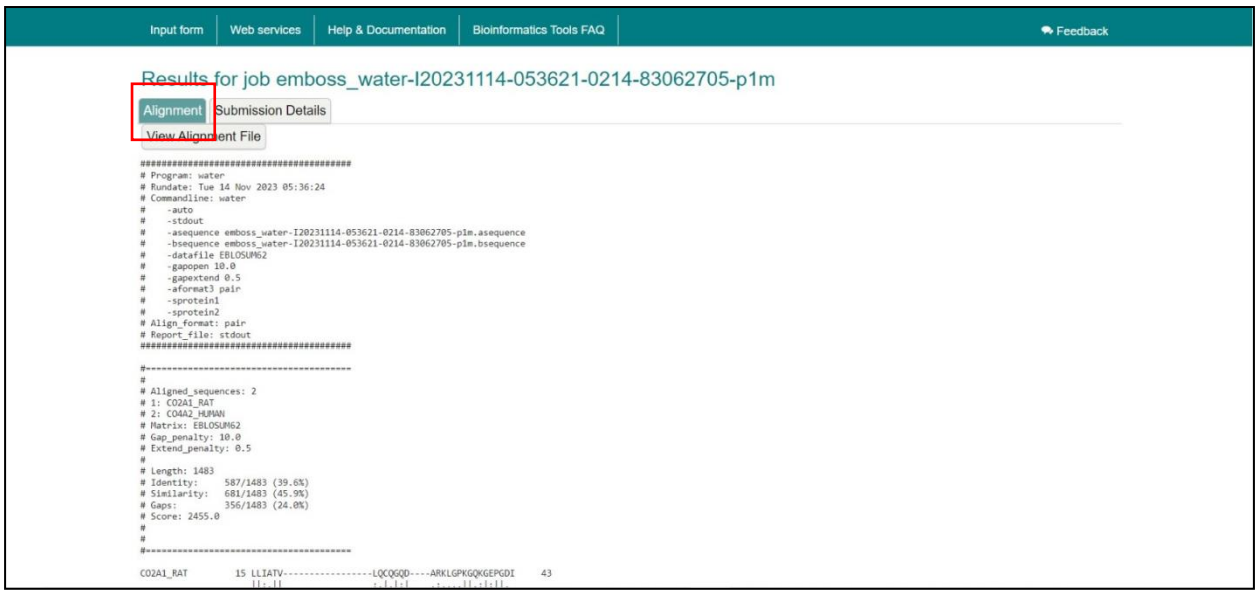


Figure 6: Results page of the submitted query with Alignment option

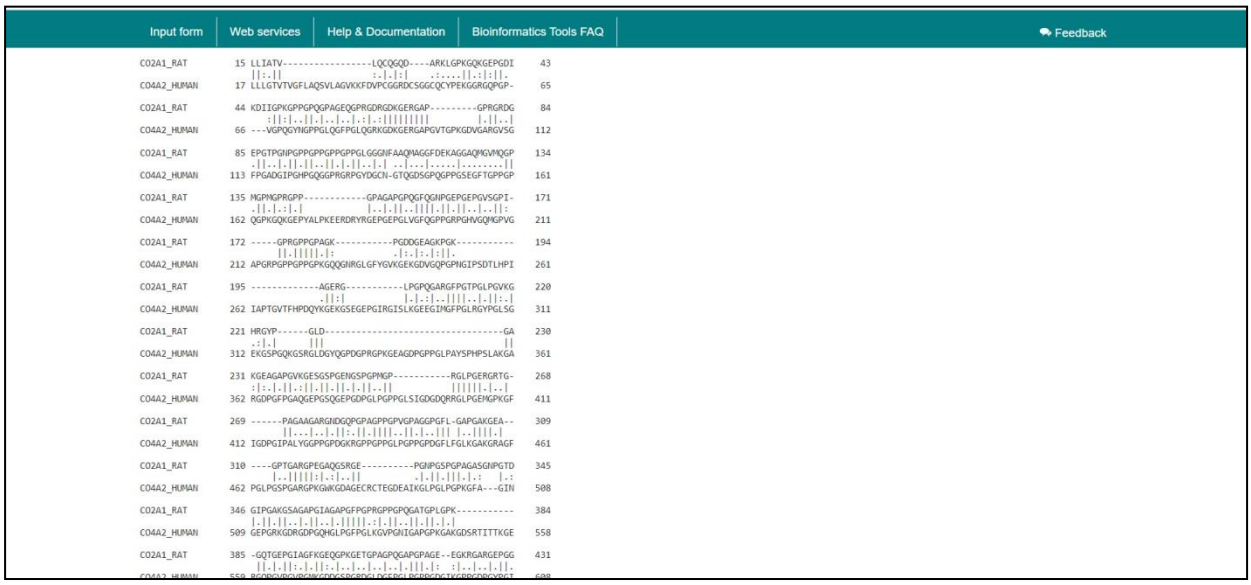


Figure 6a: Results page of the submitted query with Alignment option

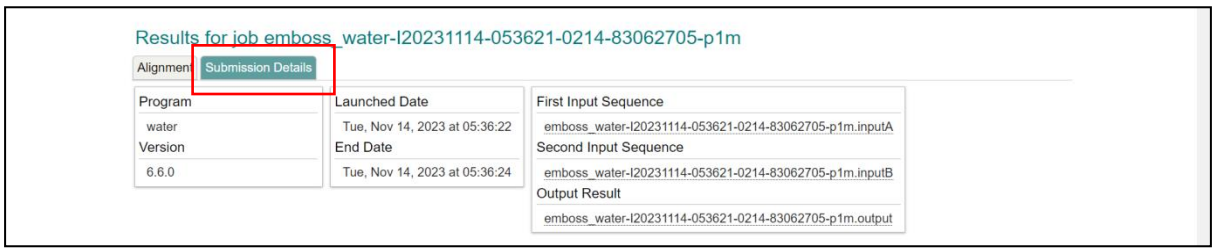


Figure 7: View of Submission details

RESULTS:

By exploring local pairwise sequence alignment using EMBOSS Water Tool, the results were observed and studied for query for protein query 'collagen' for organism *Rattus norvegicus* (UniProt ID: P05539) and *Homo sapiens* (UniProt ID: P08572) and it was observed that the local pairwise sequence alignments of the two organisms were found to be identical upon comparison, as the sequence identity is 39.6%.

Length	1483
Identity	587/14683 (39.6%)
Similarity	681/1483 (45.9%)
Gaps	356/1483 (69.3%)
Score	2455.0

CONCLUSION:

EMBOSS Water tool, for Local Pairwise Sequence Alignment, was explored by comparative study of protein collagen of two different organisms, namely, *Rattus norvegicus* (UniProt ID: P05539) and *Homo sapiens* (UniProt ID: P08572).

REFERENCES:

1. Smith TF, Waterman MS (1981) *J. Mol. Biol* 147(1). https://emboss.sourceforge.net/apps/release/6.6/emboss/apps/water.html
2. H. Jawad, R.A. Brown, in *Comprehensive Biotechnology*, 2011. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/collagen

DATE: 01/11/2023

WEBLEM 6
INTRODUCTION TO SEQUENCE ALIGNMENT TOOLS

INTRODUCTION:

Alignment of biological sequences is a fundamental task in bioinformatics. It involves identifying regions of similarity between two or more sequences, which can then be used to infer functional, structural, or evolutionary relationships. Sequence alignment is the problem of comparing biological sequences by searching for a series of nucleotides or amino acids that appear in the same order in the input sequences, possibly introducing gaps into them. When there are two sequences, it is called pairwise sequence alignment; otherwise, it is called multiple sequence alignment (MSA). Global alignment is to find the best match between the entire sequences.

Most MSA methods are based on one of the two pairwise alignment algorithms: the optimal algorithm proposed by Needleman and Wunsch (NW) for global alignment, and the improvement to the NW algorithm proposed by Smith and Waterman (SW) to obtain the local alignment. Various algorithms are employed for sequence alignment, two prominent ones being the Needleman-Wunsch algorithm and the Smith-Waterman algorithm.

The Needleman-Wunsch algorithm performs global alignment, comparing entire sequences, while the Smith-Waterman algorithm is utilized for local alignment, identifying regions of similarity within sequences. These algorithms form the backbone of sequence alignment studies and are accessible through powerful bioinformatics tools available under EMBOSS (European Molecular Biology Open Software Suite). Both algorithms are composed of three phases: initialization, distance matrix computation and trace back. Nevertheless, they differ in their applied techniques at each phase. There are many different techniques used in sequence alignment methods, such as heuristic algorithms, and dynamic programming. Although they ensure the best alignment, dynamic programming methods (such as Needleman-Wunsch and Smith-Waterman) can be computationally demanding for longer sequences. For big datasets, heuristic approaches frequently yield near-optimal alignments, by favoring optimality for of speed and efficiency.

Among the widely used tools and methods, BLAST (Basic Local Alignment Search Tool) and FASTA (Fast Alignment Search Tool) are pivotal in bioinformatics. BLAST uses heuristic methods for comparing sequences quickly and efficiently against large databases, allowing rapid identification of homologous sequences. FASTA combines heuristic methods with probability models to perform quick sequence alignments and similarity searches. These tools are used by researchers in a wide range of fields to identify homologous sequences, infer evolutionary relationships, identify functional and structural motifs, and design primers and probes.

Pairwise Alignment Tools

Pairwise alignment tools are typically used to identify regions of similarity between two sequences of unknown evolutionary relationship. They work by comparing the two sequences and identifying regions of identical or similar characters. Gaps are inserted between the

characters of the two sequences so that the identical or similar characters are aligned in successive columns.

BLAST:

BLAST (Basic Local Alignment Search Tool) is a family of sequence alignment algorithms and programs designed to search for regions of similarity between biological sequences. It is used to search for homologous sequences in a database of known sequences, which can be used to identify genes, infer evolutionary relationships, and design primers and probes. It works by comparing a query sequence to a database of sequences using a heuristic approach. This means that it does not search the entire database for matches, but instead uses a number of shortcuts to identify potential matches. The first step in BLAST is to break the query sequence into short segments, called words. The length of the words depends on the type of sequence being searched (e.g., DNA or protein). BLAST then searches the database for sequences that contain the same words as the query sequence. If a match is found, BLAST extends the alignment in both directions to find the longest possible alignment. BLAST calculates a score for each alignment, which is based on the similarity of the two sequences and the presence of gaps. The higher the score, the more similar the two sequences are. BLAST then reports the alignments with the highest scores.

Types of BLAST:

There are five types (variants) of BLAST that are differentiated based on the type of sequence (DNA or protein) of the query and database sequences.

1. **BLASTN** compares a nucleotide query sequence to a nucleotide sequence database.
2. **BLASTP** compares a protein query sequence to a protein sequence database.
3. **BLASTX** compares a nucleotide query sequence to a protein sequence database by translating the query sequence into its six possible reading frames and aligning them with the protein sequences.
4. **TBLASTN** compares a protein query sequence to a nucleotide sequence database by translating the nucleotide sequences in all six reading frames and aligning them with the protein sequence.
5. **TBLASTX** compares a nucleotide query sequence to a nucleotide sequence database by translating the query sequence in all six reading frames and aligning them with the nucleotide sequences.

FASTA:

FASTA (Fast Alignment Search Tool) is a sequence alignment algorithm and program that is used to search for regions of similarity between biological sequences. It works by first building a hash table of the query sequence. The hash table is a data structure that allows FASTA to quickly find all of the sequences in the database that contain the same words as the query sequence. It then aligns the query sequence to each of the matching sequences in the database to find the longest possible alignment. It calculates a score for each alignment, which is based on the similarity of the two sequences and the presence of gaps. The higher the score, the more similar the two sequences are. It then reports the alignments with the highest scores. It is often used in conjunction with BLAST to identify and analyze homologous sequences. FASTA is

also used to design primers and probes for PCR and other molecular biology techniques.

PSI-BLAST:

PSI-BLAST (Position-Specific Iterative BLAST) is a sequence alignment tool that uses a position-specific scoring matrix (PSSM) to search for distant homologs in protein sequences. It is particularly well-suited for identifying homologs that have diverged significantly from their known relatives. It works by first running a regular BLAST search of the protein sequence database using the query sequence. This produces a list of initial hits. It then constructs a PSSM from the alignments of the initial hits. The PSSM is a statistical model that describes the probability of each amino acid at each position in the alignment. PSI-BLAST then uses the PSSM to search the protein sequence database again. This produces a list of new hits. It then repeats this process, using the PSSM from the previous iteration to search for new hits. PSI-BLAST continues to iterate until the PSSM no longer changes or until a certain number of iterations have been reached. PSI-BLAST then reports the alignments with the highest scores.

PHI-BLAST:

PHI-BLAST (Phylogenetically Inconsistent BLAST) is a sequence alignment tool that uses a probabilistic model to search for distant homologs in protein sequences. It is particularly well-suited for identifying homologs that have diverged significantly from their known relatives. It works by first building a phylogenetic tree of the known homologs of the query sequence. It then uses this tree to generate a position-specific scoring matrix (PSSM) for each node in the tree. The PSSM is a statistical model that describes the probability of each amino acid at each position in the alignment. It then searches the database of protein sequences for sequences that match the PSSMs at the nodes of the phylogenetic tree. It does this by calculating a score for each alignment based on the similarity of the sequences and the PSSM. The higher the score, the more similar the sequences are and the more likely they are to be homologous. It then reports the alignments with the highest scores. It also reports the probability that each alignment is a true homolog. This probability is based on the score of the alignment, the PSSM of the node in the phylogenetic tree, and the phylogenetic relationships between the sequences in the alignment. PHI-BLAST is a powerful tool for identifying distant homologs. It is used by researchers in a wide range of fields, including genetics, genomics, proteomics, and molecular biology.

EMBOSS Needle:

EMBOSS Needle is a pairwise sequence alignment tool that uses the Needleman- Wunsch algorithm to produce global alignments. A global alignment is an alignment that aligns the entire length of both sequences. It works by comparing the two sequences and identifying regions of identical or similar characters. Gaps are inserted between the characters of the two sequences so that the identical or similar characters are aligned in successive columns. It calculates a score for each alignment, which is based on the similarity of the two sequences and the presence of gaps. The higher the score, the more similar the two sequences are. It then reports the alignment with the highest score. EMBOSS Needle is a powerful tool for aligning biological sequences and it is particularly well-suited for aligning sequences of known evolutionary relationship or sequences with low levels of divergence.

EMBOSS Water:

EMBOSS Water is a pairwise alignment tool that uses the Smith-Waterman algorithm to produce local alignments. This means that only the most similar regions of the two sequences are aligned. It is a good choice for aligning sequences of unknown evolutionary relationship or sequences with high levels of divergence. It works by comparing the two sequences and identifying regions of identical or similar characters. Gaps are inserted between the characters of the two sequences so that the identical or similar characters are aligned in successive columns. It then calculates a score for each alignment, which is based on the similarity of the two sequences and the presence of gaps. The higher the score, the more similar the two sequences are. It then reports the alignment with the highest score. It is a powerful tool for aligning biological sequences. It is often used in conjunction with other alignment tools, such as BLAST and FASTA, to identify and analyze homologous sequences. EMBOSS Water is also used to design primers and probes for PCR and other molecular biology techniques.

REFERENCES:

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2
2. Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4
3. Bhagwat, M., & Aravind, L. (2007). PSI-BLAST Tutorial. In Methods in molecular biology (pp. 177–186). https://doi.org/10.1007/978-1-59745-514-5_10
4. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., López, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, 7(1). https://doi.org/10.1038/msb.2011.75

DATE: 01/11/2023

WEBLEM 6(A)
BASIC LOCAL ALIGNMENT SEARCH TOOL (BLAST)
(URL: https://blast.ncbi.nlm.nih.gov)

AIM:

To study and explore similar sequences of the protein albumin (UniProt ID: P02768) by using Basic Local Alignment Search Tool (BLAST).

INTRODUCTION:

BLAST (Basic Local Alignment Search Tool) is an algorithm and program for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA and/or RNA sequences. A BLAST search enables a researcher to compare a subject protein or nucleotide sequence (called a query) with a library or database of sequences, and identify database sequences that resemble the query sequence above a certain threshold. BLAST (Basic Local Alignment Search Tool) has become the defacto standard in search and alignment tools [Altschul et al., 1990]. The BLAST algorithm works by finding a short, or local, region of high similarity between two sequences, and then extending this match out from this starting point to both the left and the right. A score is assigned to the match. The score will increase as more residues are found to match and will decrease if there are gaps in the alignment. Alignments with a score that exceeds a certain threshold are reported in the output.

BLAST searches for high scoring sequence alignments between the query sequence and the existing sequences in the database using a heuristic approach that approximates the Smith-Waterman algorithm.

BLAST tool can be used to identify unknown sequences by comparing them with known sequences in a database which helps in predicting the functions of proteins or genes which can be used in phylogenetic analysis as well as in identifying functionally conserved domains within proteins which is important for predicting the functions of proteins.

Albumin:

Albumin is a family of globular proteins, with the most common members being the serum albumins. All proteins within the albumin family are water-soluble, moderately soluble in concentrated salt solutions, and susceptible to heat denaturation. Albumins are commonly present in blood plasma and distinguish themselves from other blood proteins by their lack of glycosylation. Compounds containing albumins are termed albuminoids. Several blood transport proteins share an evolutionary relationship within the albumin family, including serum albumin, alpha-fetoprotein, vitamin D-binding protein, and afamin. This family is exclusively found in vertebrates. In a broader sense, the term "albumins" may refer to other proteins that coagulate under specific conditions.

METHODOLOGY:

1. Open the Homepage of the UniProt database and search for the query of Albumin protein.
2. Select one entry from the results for *Homo sapiens* (UniProt ID: P02768) and download its FASTA sequence in canonical format.
3. Open the homepage of BLAST and select Protein BLAST, i.e., BLASTP.
4. Paste the FASTA sequence in 'Enter Query Sequence' box.
5. Set the desired parameters.
6. Run the BLAST.

OBSERVATIONS:

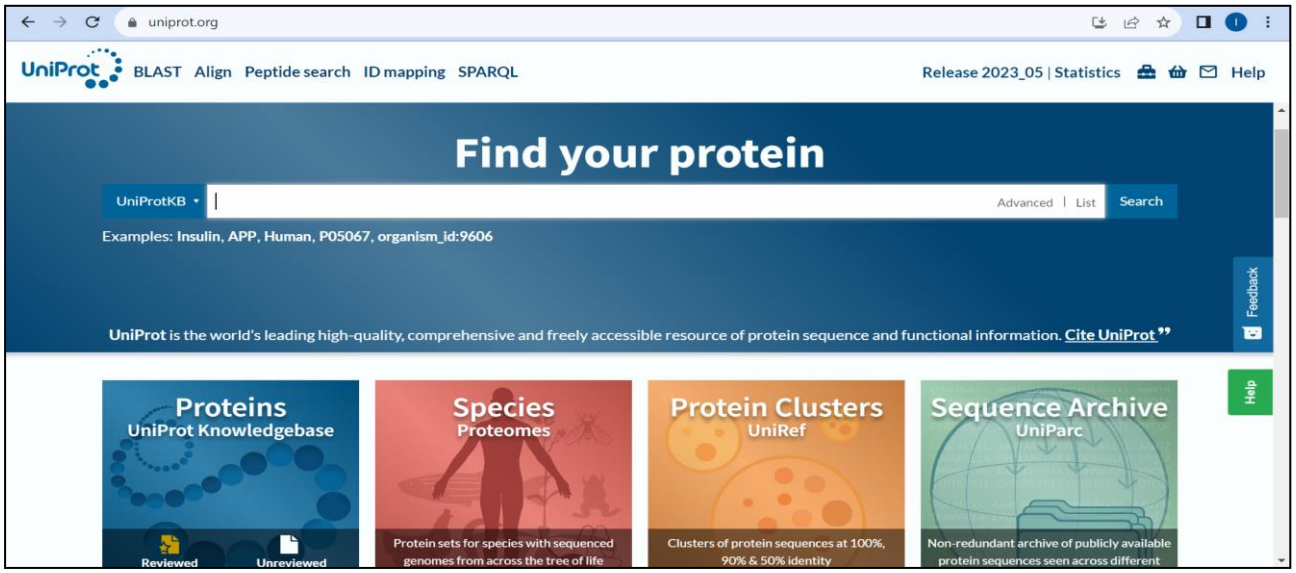


Figure 1: Homepage of the UniProt Database

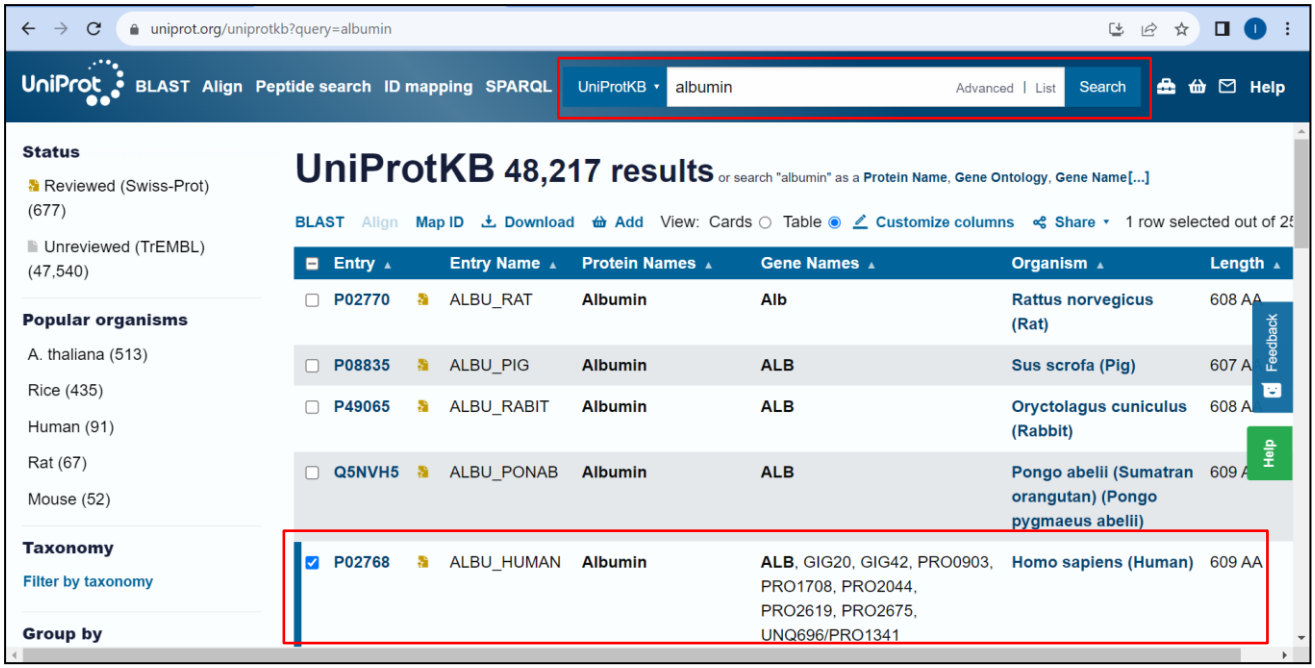


Figure 2: Searching for the query albumin and selecting (UniProt ID: P02768)

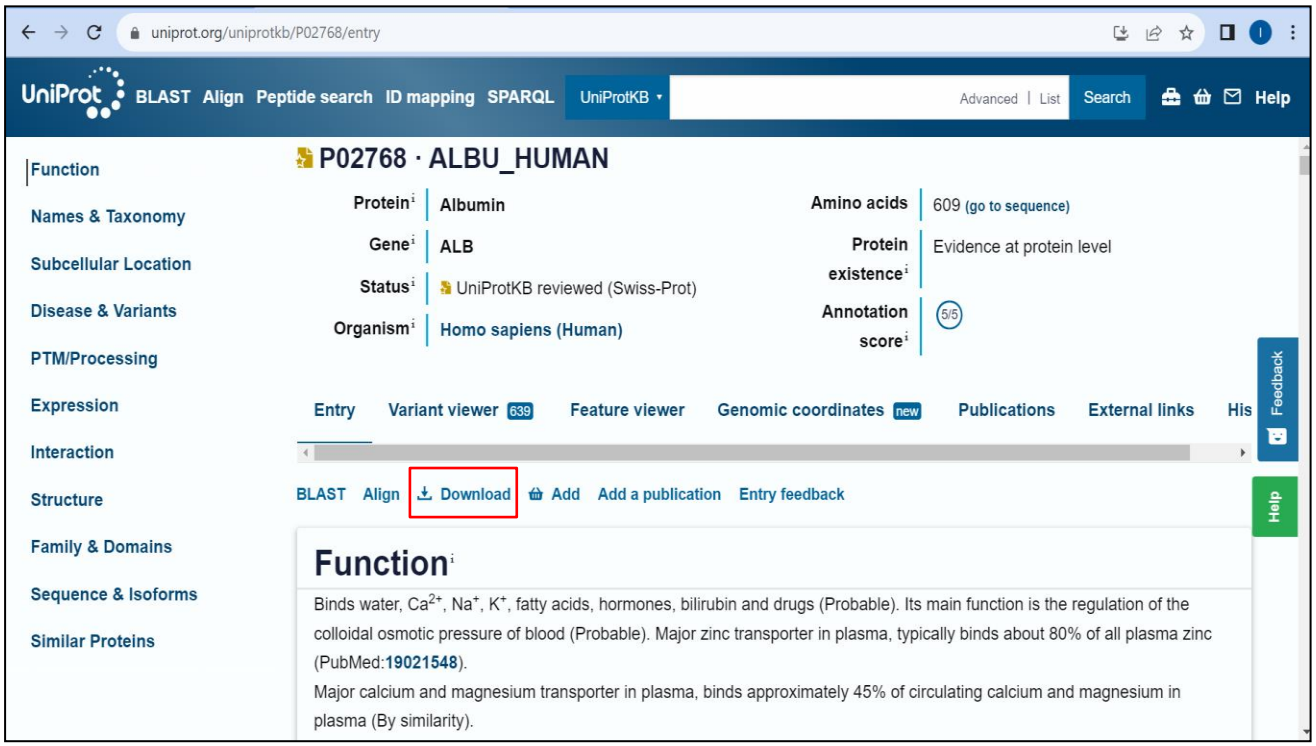


Figure 3: Download option for retrieving FASTA sequence

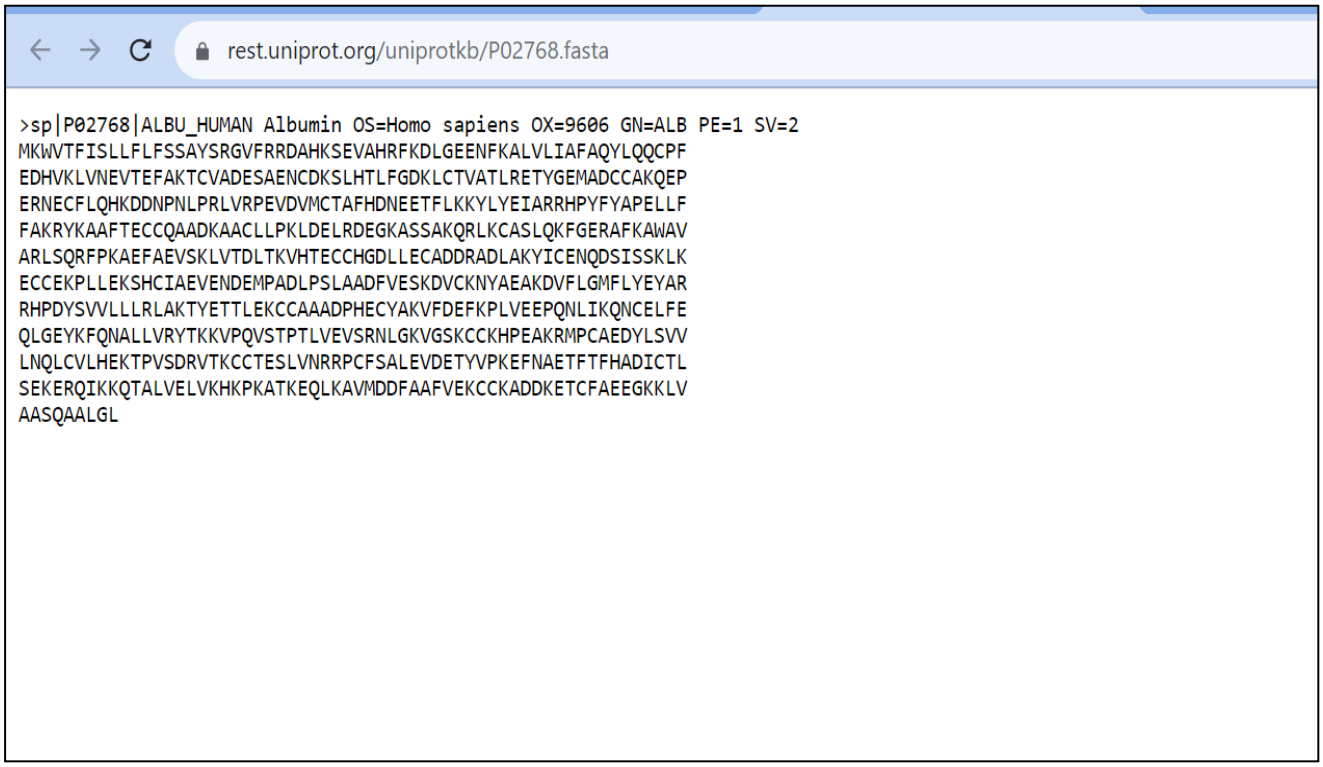


Figure 4: FASTA sequence in canonical format

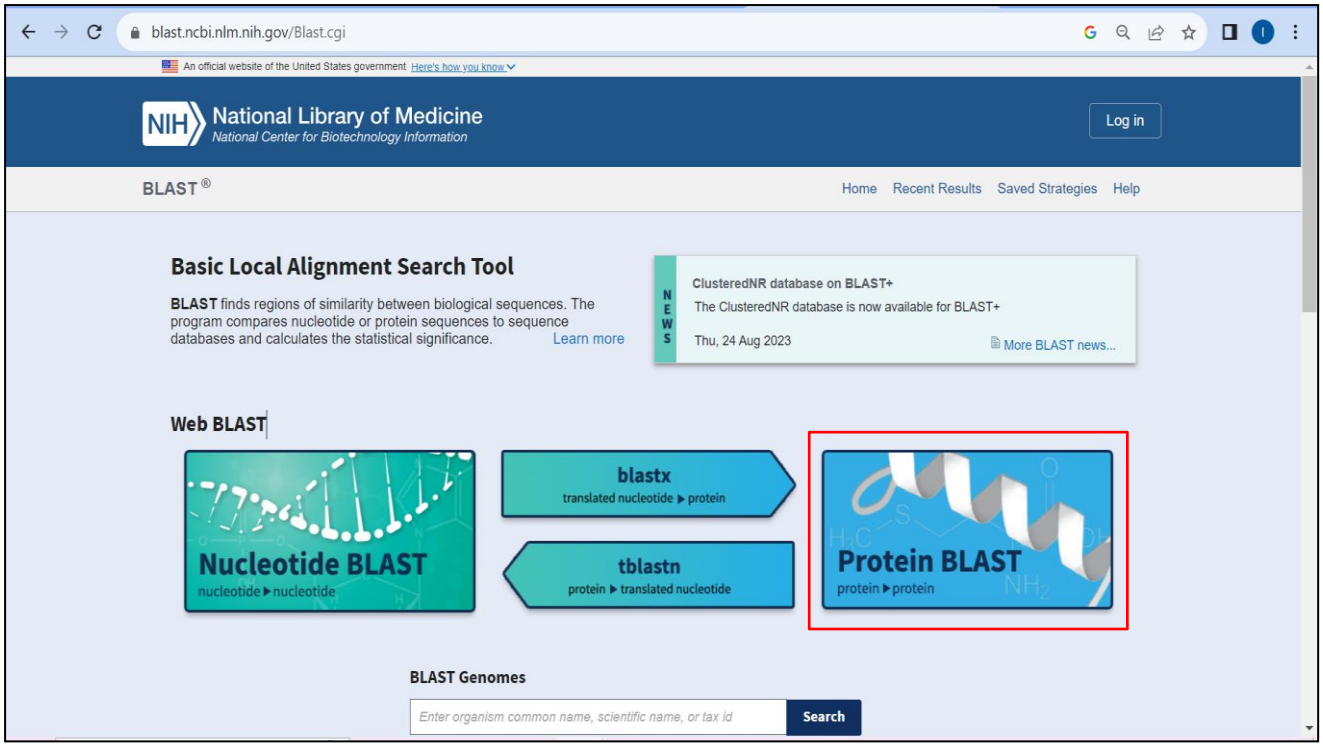


Figure 5: Homepage of Basic Local Alignment Search Tool (BLAST)

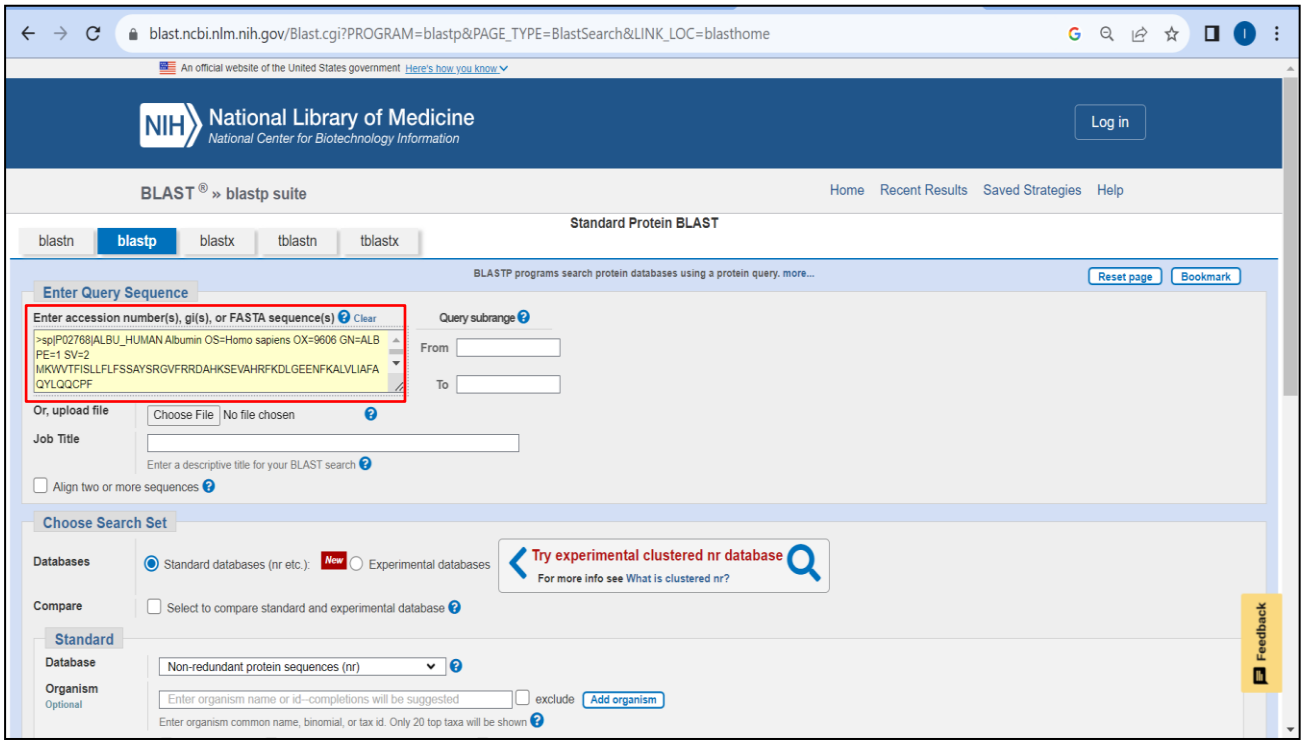


Figure 6: FASTA sequence pasted in 'Enter Query Sequence' box

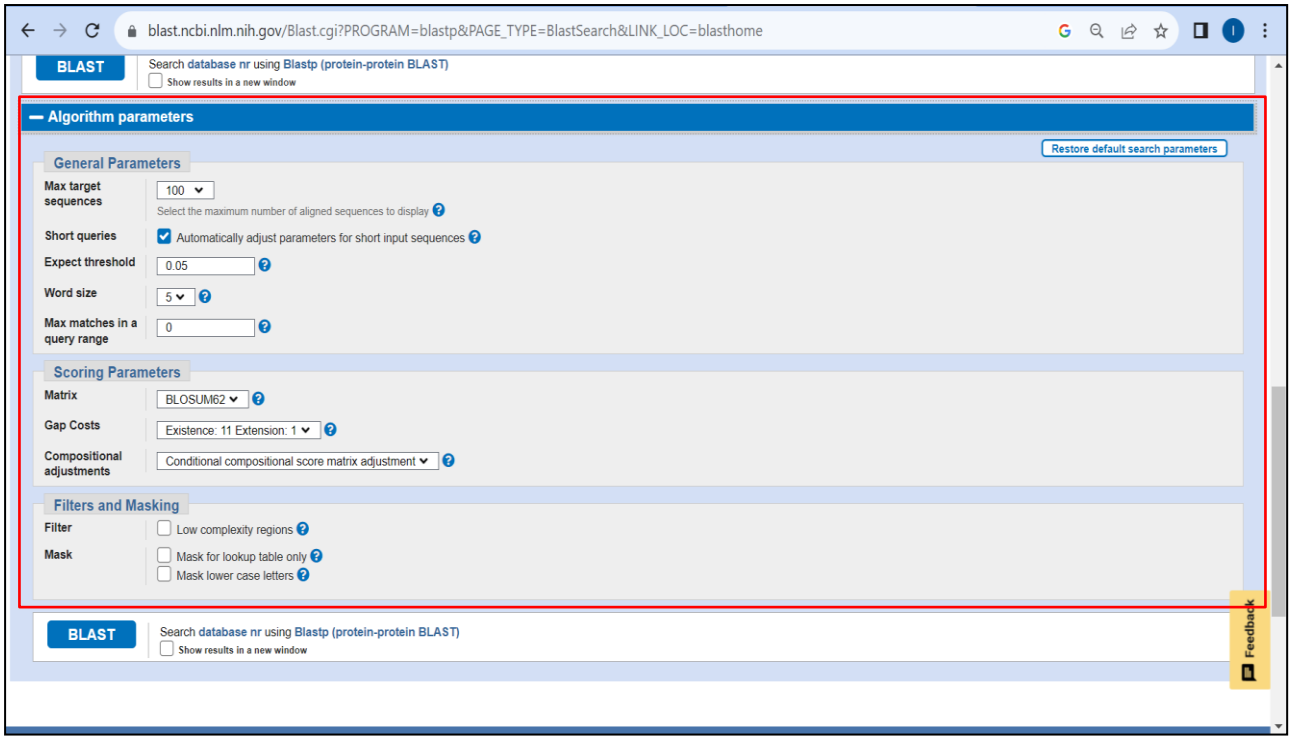


Figure 7: Setting the Algorithm parameters

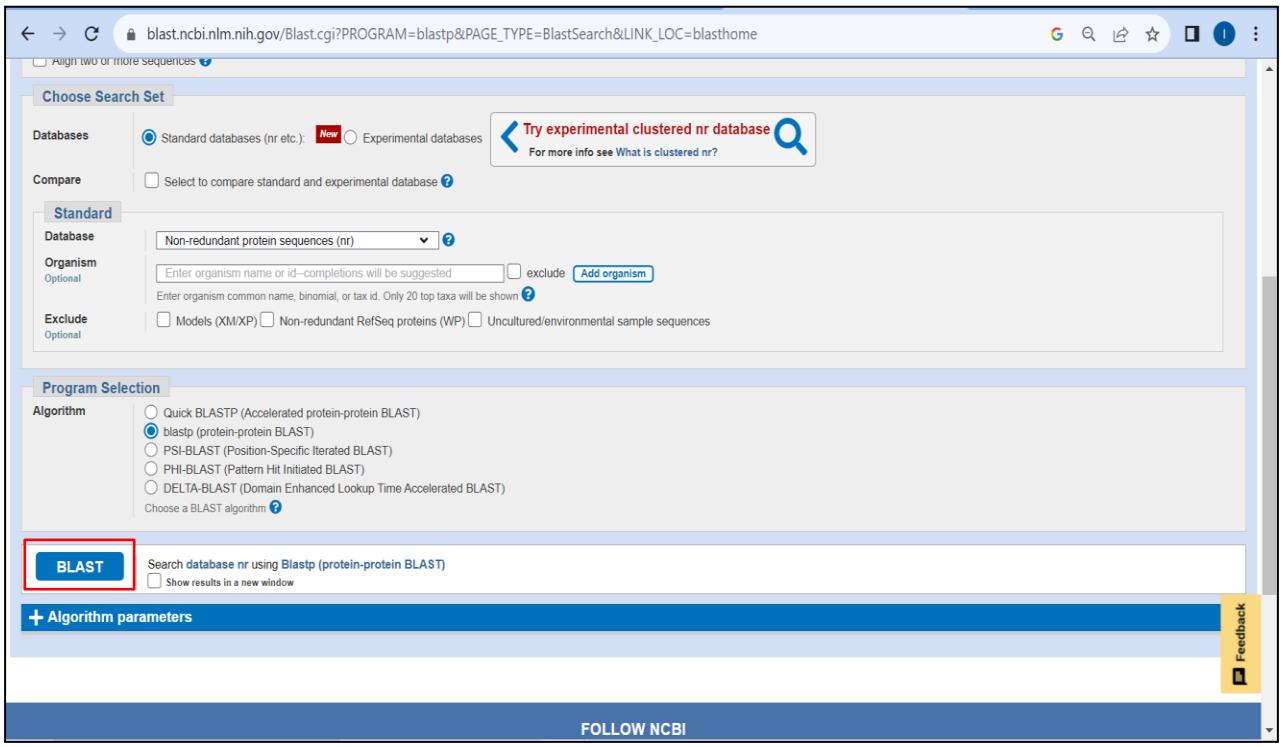


Figure 8: Running BLAST

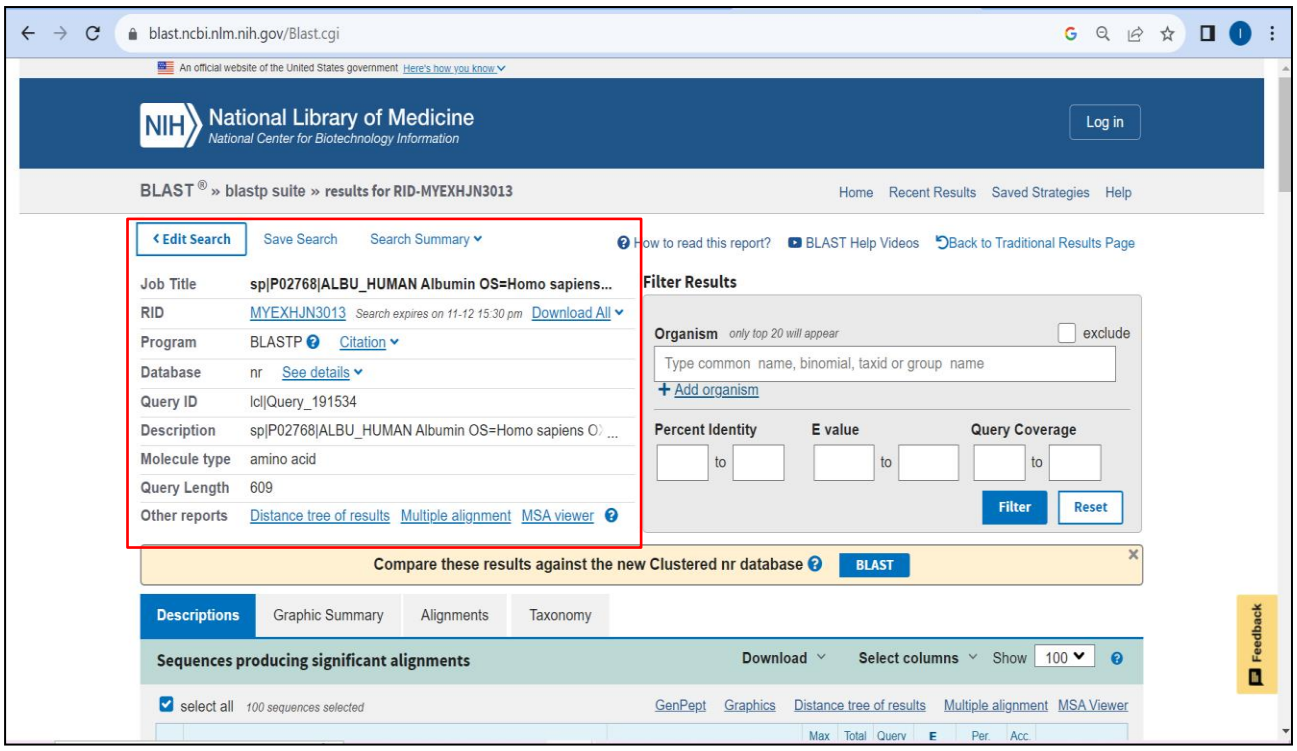
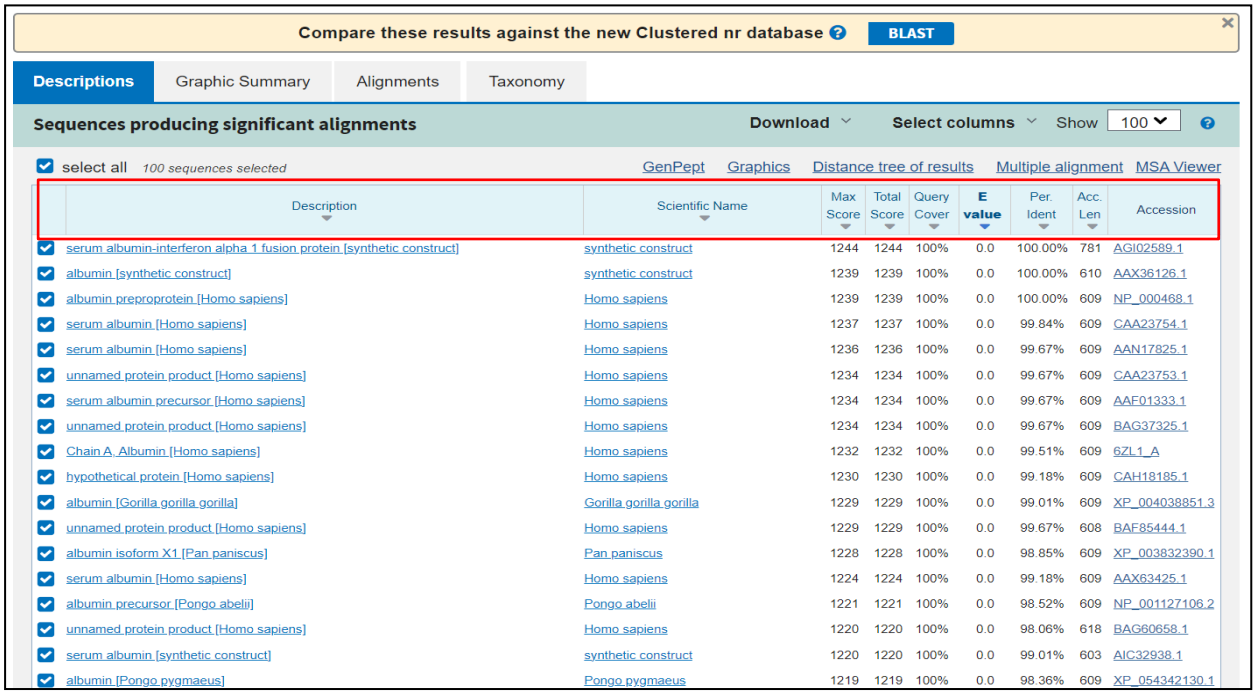


Figure 9: Results for the query, Header Section (UniProt ID: P02768)



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
serum albumin-interferon alpha 1 fusion protein [synthetic construct]	synthetic construct	1244	1244	100%	0.0	100.00%	781	AGI02589.1
albumin [synthetic construct]	synthetic construct	1239	1239	100%	0.0	100.00%	610	AAX36126.1
albumin preproprotein [Homo sapiens]	Homo sapiens	1239	1239	100%	0.0	100.00%	609	NP_000468.1
serum albumin [Homo sapiens]	Homo sapiens	1237	1237	100%	0.0	99.84%	609	CAA23754.1
serum albumin [Homo sapiens]	Homo sapiens	1236	1236	100%	0.0	99.67%	609	AAN17825.1
unnamed protein product [Homo sapiens]	Homo sapiens	1234	1234	100%	0.0	99.67%	609	CAA23753.1
serum albumin precursor [Homo sapiens]	Homo sapiens	1234	1234	100%	0.0	99.67%	609	AAF01333.1
unnamed protein product [Homo sapiens]	Homo sapiens	1234	1234	100%	0.0	99.67%	609	BAG37325.1
Chain A, Albumin [Homo sapiens]	Homo sapiens	1232	1232	100%	0.0	99.51%	609	6ZL1_A
hypothetical protein [Homo sapiens]	Homo sapiens	1230	1230	100%	0.0	99.18%	609	CAH18185.1
albumin [Gorilla gorilla gorilla]	Gorilla gorilla gorilla	1229	1229	100%	0.0	99.01%	609	XP_004038851.3
unnamed protein product [Homo sapiens]	Homo sapiens	1229	1229	100%	0.0	99.67%	608	BAF85444.1
albumin isoform X1 [Pan paniscus]	Pan paniscus	1228	1228	100%	0.0	98.85%	609	XP_003832390.1
serum albumin [Homo sapiens]	Homo sapiens	1224	1224	100%	0.0	99.18%	609	AAX63425.1
albumin precursor [Pongo abelii]	Pongo abelii	1221	1221	100%	0.0	98.52%	609	NP_001127106.2
unnamed protein product [Homo sapiens]	Homo sapiens	1220	1220	100%	0.0	98.06%	618	BAG60658.1
serum albumin [synthetic construct]	synthetic construct	1220	1220	100%	0.0	99.01%	603	AIC32938.1
albumin [Pongo pygmaeus]	Pongo pygmaeus	1219	1219	100%	0.0	98.36%	609	XP_054342130.1

Figure 10: Result for Description Section

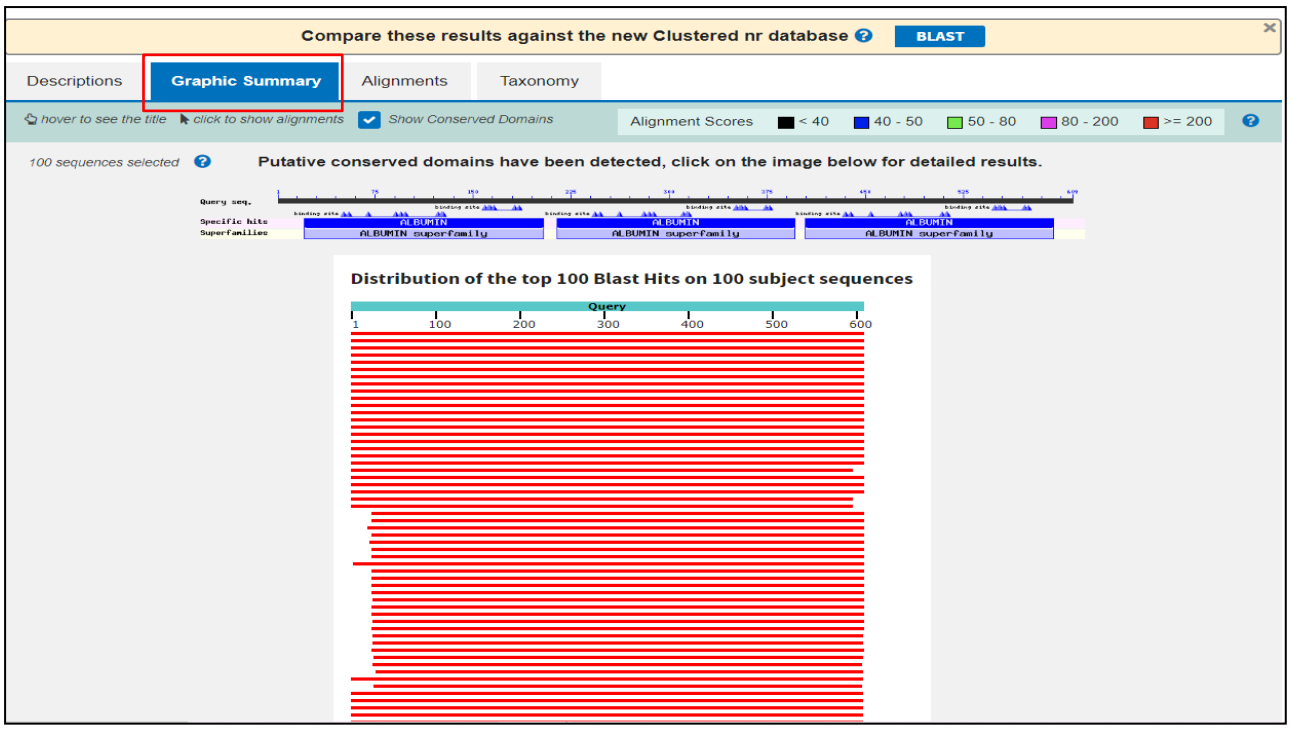


Figure 11: Result for Graphic Summary Section

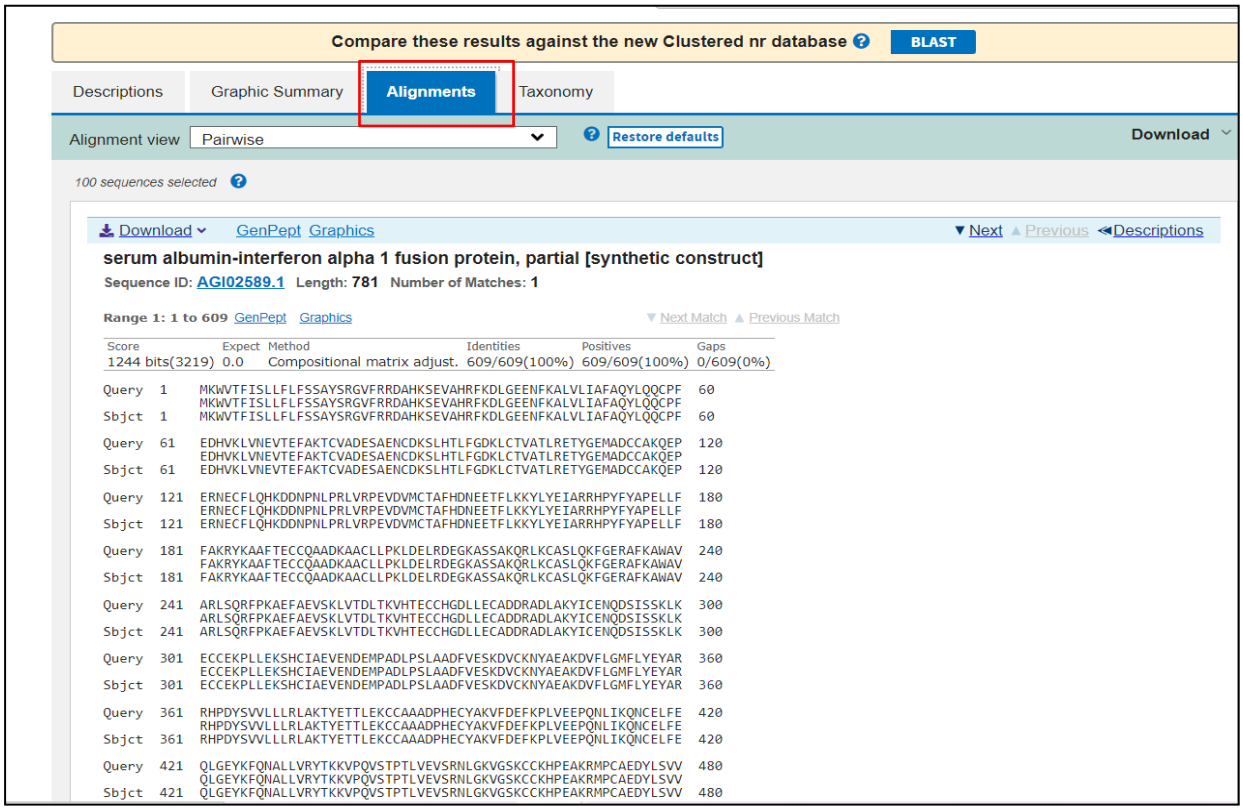
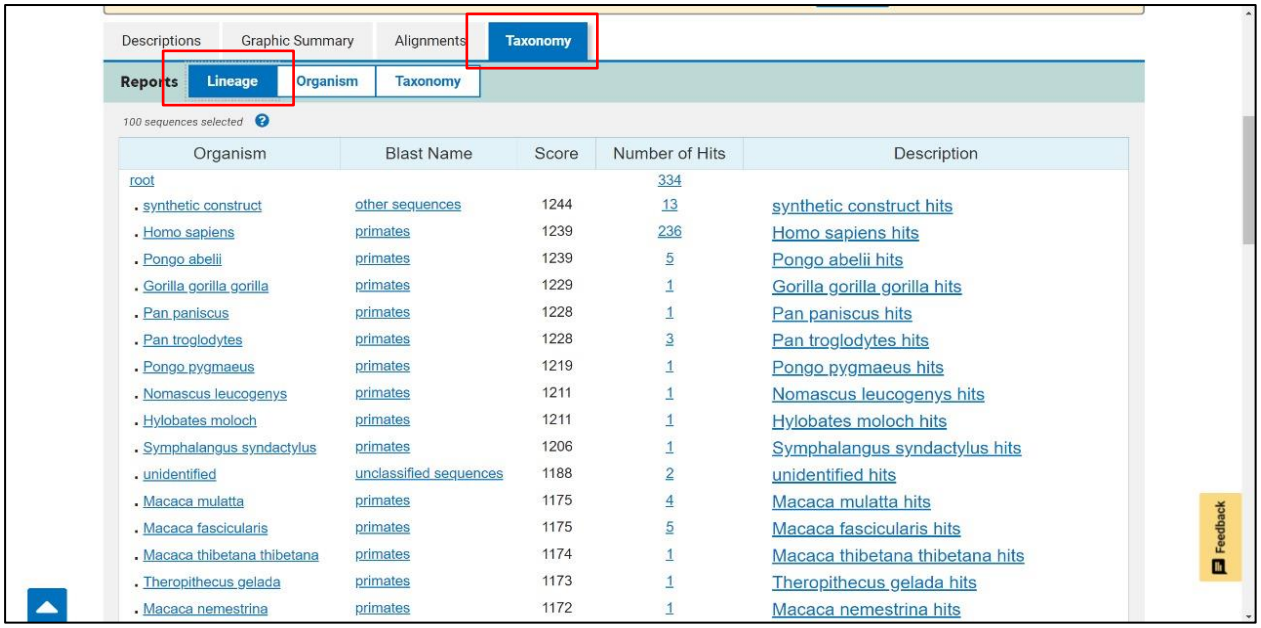


Figure 12: Result for Alignment Section



Descriptions | Graphic Summary | Alignments | Taxonomy

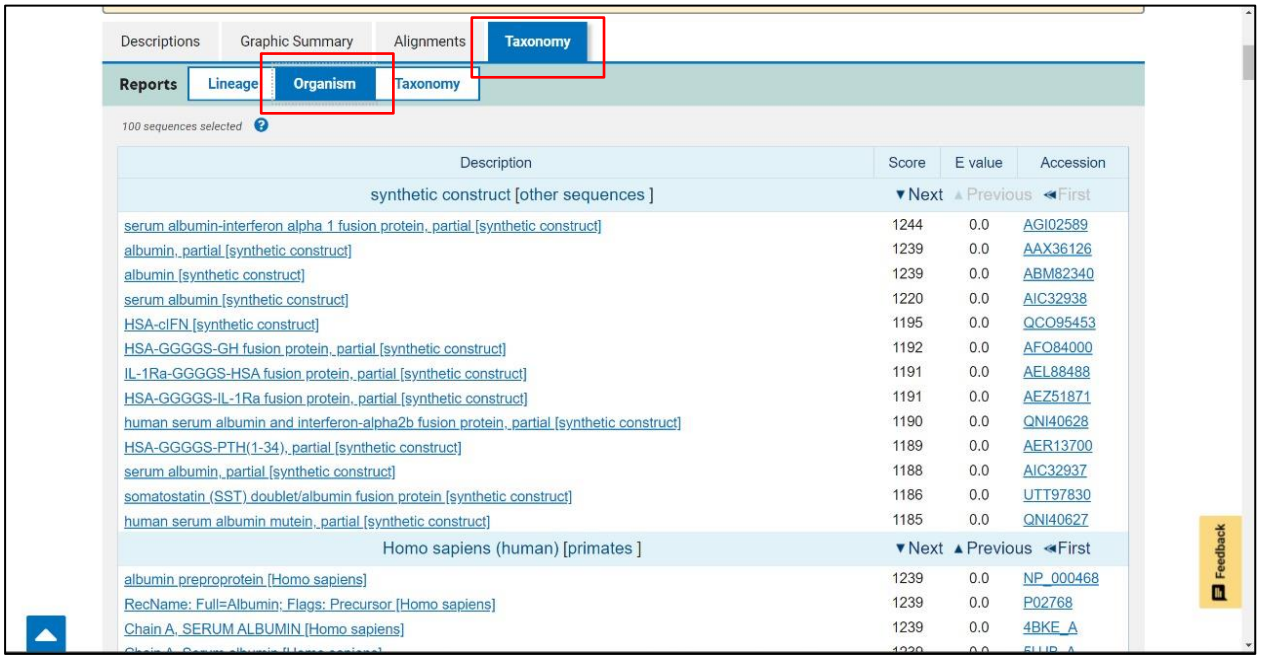
Reports | Lineage | Organism | Taxonomy

100 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
root			334	
synthetic construct	other sequences	1244	13	synthetic construct hits
Homo sapiens	primates	1239	236	Homo sapiens hits
Pongo abelii	primates	1239	5	Pongo abelii hits
Gorilla gorilla gorilla	primates	1229	1	Gorilla gorilla gorilla hits
Pan paniscus	primates	1228	1	Pan paniscus hits
Pan troglodytes	primates	1228	3	Pan troglodytes hits
Pongo pygmaeus	primates	1219	1	Pongo pygmaeus hits
Nomascus leucogenys	primates	1211	1	Nomascus leucogenys hits
Hylobates moloch	primates	1211	1	Hylobates moloch hits
Symphalangus syndactylus	primates	1206	1	Symphalangus syndactylus hits
unidentified	unclassified sequences	1188	2	unidentified hits
Macaca mulatta	primates	1175	4	Macaca mulatta hits
Macaca fascicularis	primates	1175	5	Macaca fascicularis hits
Macaca thibetana thibetana	primates	1174	1	Macaca thibetana thibetana hits
Theropithecus gelada	primates	1173	1	Theropithecus gelada hits
Macaca nemestrina	primates	1172	1	Macaca nemestrina hits

Feedback

Figure 13: Result for Taxonomy Section based on Lineage



Descriptions | Graphic Summary | Alignments | Taxonomy

Reports | Lineage | Organism | Taxonomy

100 sequences selected

Description	Score	E value	Accession
synthetic construct [other sequences]	▼Next	▲Previous	◀◀First
serum albumin-interferon alpha 1 fusion protein, partial [synthetic construct]	1244	0.0	AGI02589
albumin, partial [synthetic construct]	1239	0.0	AAX36126
albumin [synthetic construct]	1239	0.0	ABM82340
serum albumin [synthetic construct]	1220	0.0	AIC32938
HSA-cIFN [synthetic construct]	1195	0.0	QCO95453
HSA-GGGGS-GH fusion protein, partial [synthetic construct]	1192	0.0	AFO84000
IL-1Ra-GGGGS-HSA fusion protein, partial [synthetic construct]	1191	0.0	AEL88488
HSA-GGGGS-IL-1Ra fusion protein, partial [synthetic construct]	1191	0.0	AEZ51871
human serum albumin and interferon-alpha2b fusion protein, partial [synthetic construct]	1190	0.0	QNI40628
HSA-GGGGS-PTH(1-34), partial [synthetic construct]	1189	0.0	AER13700
serum albumin, partial [synthetic construct]	1188	0.0	AIC32937
somatostatin (SST) doublet/albumin fusion protein [synthetic construct]	1186	0.0	UTT97830
human serum albumin mutein, partial [synthetic construct]	1185	0.0	QNI40627
Homo sapiens (human) [primates]	▼Next	▲Previous	◀◀First
albumin preproprotein [Homo sapiens]	1239	0.0	NP_000468
RecName: Full=Albumin; Flags: Precursor [Homo sapiens]	1239	0.0	P02768
Chain A, SERUM ALBUMIN [Homo sapiens]	1239	0.0	4BKE_A

Feedback

Figure 13a: Result for Taxonomy Section based on Organism

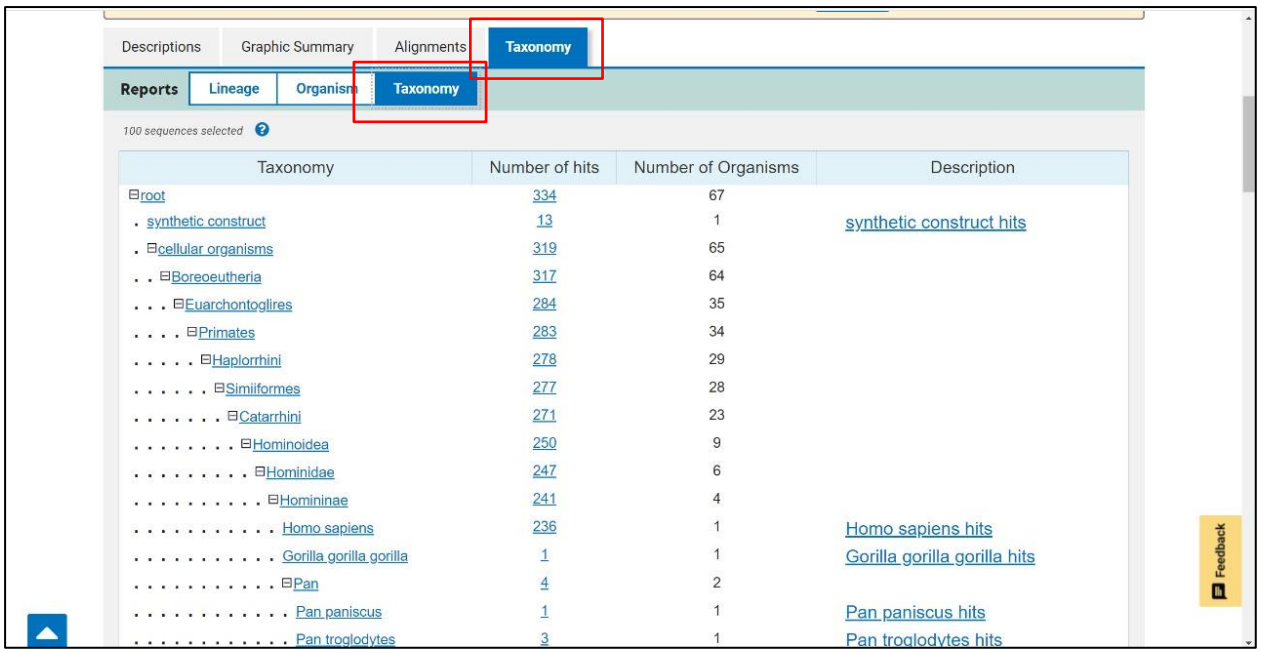


Figure 13b: Result for Taxonomy Section based on Taxonomy

RESULTS:

The Basic Local Alignment Search Tool (BLAST) was used to explore the protein sequences similar to the protein sequence of albumin (UniProt ID: P02768). The query sequence is found 100% identical to three sequence entries.

Sequence Title	Organism	Max Score	Total Score	E Value	Percentage Identity	Accession ID
serum albumin-interferon alpha 1 fusion protein	Synthetic construct	1244	1244	0.0	100.0%	AGI02589.1
albumin	Synthetic construct	1239	1239	0.0	100.0%	AAX36126.1
albumin preproprotein	*Homo Sapiens*	1239	1239	0.0	100.0%	NP_000468.1

CONCLUSION:

The protein sequences similar to the protein sequence of albumin (UniProt ID: P02768) were studied by exploring the Basic Local Alignment Search Tool (BLAST).

REFERENCES:

1. Xiong, J. (2006). *Essential Bioinformatics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511806087
2. S. Sugio, A. Kashima, S. Mochizuki, M. Noda, K. Kobayashi, Crystal structure of human serum albumin at 2.5 Å resolution, *Protein Engineering, Design and Selection*, Volume 12, Issue 6, June 1999, Pages 439–446, https://doi.org/10.1093/protein/12.6.439

3. He, X., Carter, D. Atomic structure and chemistry of human serum albumin. *Nature* 358, 209–215 (1992). https://doi.org/10.1038/358209a0

DATE: 01/11/2023

WEBLEM 6(B)
FASTA TOOL
(URL: https://www.ebi.ac.uk/Tools/sss/fasta/)

AIM:

To study protein sequence similarity by exploring FASTA tool for the query maltose (UniProt ID: P68187).

INTRODUCTION:

FASTA tool was originally developed for comparing protein sequences. FASTA is a text-based format for representing nucleotide or amino acid sequences. It's used in bioinformatics and biochemistry. FASTA is an abbreviation for "Fast-All". FASTA is a sequence alignment tool that takes nucleotide or protein sequences as input and compares it with existing databases. It was the first database similarity search tool developed, preceding the development of BLAST. The FASTA format allows for sequence names and comments to precede the sequences. Nucleotides or amino acids are represented using single-letter codes. For example, A => adenosine, C => cytidine, G => guanine, T => thymidine, and N => A/G/C/T (any). The original program was referred to as FASTP. It quickly became a popular tool for sequence alignment and database searching. The program has been continually updated and improved.

There are now different FASTA programs available, each used for different types of sequence searches:

1. **FASTA** compares a DNA query sequence against a database of DNA sequences or a protein query sequence against a database of protein sequences using the FASTA algorithm.
2. **SSEARCH** performs protein-protein or DNA-DNA comparisons using the SmithWaterman algorithm.
3. **GGSEARCH/GLSEARCH** works using a global alignment algorithm (GGSEARCH) or a combination of global and local alignment algorithms (GLSEARCH) to compare protein and nucleotide sequences.
4. **FASTX/FASTY** compares a DNA sequence and a database of protein sequences by translating the DNA sequence into three frames and allowing gaps and frameshifts.
5. **TFASTX/TFASTY** compares a protein sequence and a database of DNA sequences. The DNA sequence is translated in six frames – three in the forward direction and three in the reverse direction.
6. **FASTF/TFASTF** compares mixed peptide sequences against a protein (FASTF) or translated DNA (TFASTF) databases.
7. **FASTS/TFASTS** compares a set of short peptide fragments against the protein (FASTS) or translated DNA (TFASTS) databases.

1. **<u>How FASTA Works</u>**
   FASTA works by comparing a query sequence to a database of sequences to identify similar matches. The program uses a heuristic algorithm to quickly search the database and identify the most significant matches.

2. **<u>The working mechanism of FASTA is described in the following steps:</u>**
   Step 1: Identifying Regions
   The first step is identifying regions with high similarity by creating a lookup table for the query sequence. This step is also called hashing step. To create the lookup table, the query sequence is first broken down into smaller words known as k-tuples (ktup).

   Step 2: Re-Scoring
   In the second step, the ten best diagonals are rescored using suitable scoring matrices. For protein, BLOSUM50 or PAM matrix is used; for DNA sequences, the identity matrix is used. A subregion with the highest score is identified for each of the rescanned diagonal regions.

   Step 3: Joining Threshold
   Next, a score cutoff or the joining threshold is applied that excludes segments unlikely to be part of the final alignment. The library sequences are ranked based on their Initial scores.

   Step 4: Final Alignment
   Finally, the gapped alignment is refined to produce the final alignment. This is done by using the banded Smith-Waterman algorithm, which is a dynamic programming algorithm that calculates the optimal score (opt) for alignment.

<u>Maltose:</u>
Maltose-binding protein (MBP) is a part of the maltose/maltodextrin system of Escherichia coli, which is responsible for the uptake and efficient catabolism of maltodextrins. It is a complex regulatory and transport system involving many proteins and protein complexes. MBP has an approximate molecular mass of 42.5 kilodaltons.

<u>METHODOLOGY:</u>

1. The protein FASTA (canonical) sequence for the desired protein for the query of ‘Maltose’ (UniProt ID: P68187) was retrieved from the UniProt Database.
2. Open the homepage of EBI – FASTA tool. Select the desired Protein Database and paste the retrieved FASTA (canonical) sequence of Maltose (UniProt ID: P68187) in the query box of the EBI – FASTA tool.
3. Set the desired parameters and select the ‘SUBMIT’ option to submit the query to the tool.
4. The results were shown in different tabs, namely, Submission Information, Tool Output, Graphic Output, Functional Forecasts, and Summary Table.
5. Interpret the results obtained.

OBSERVATIONS:

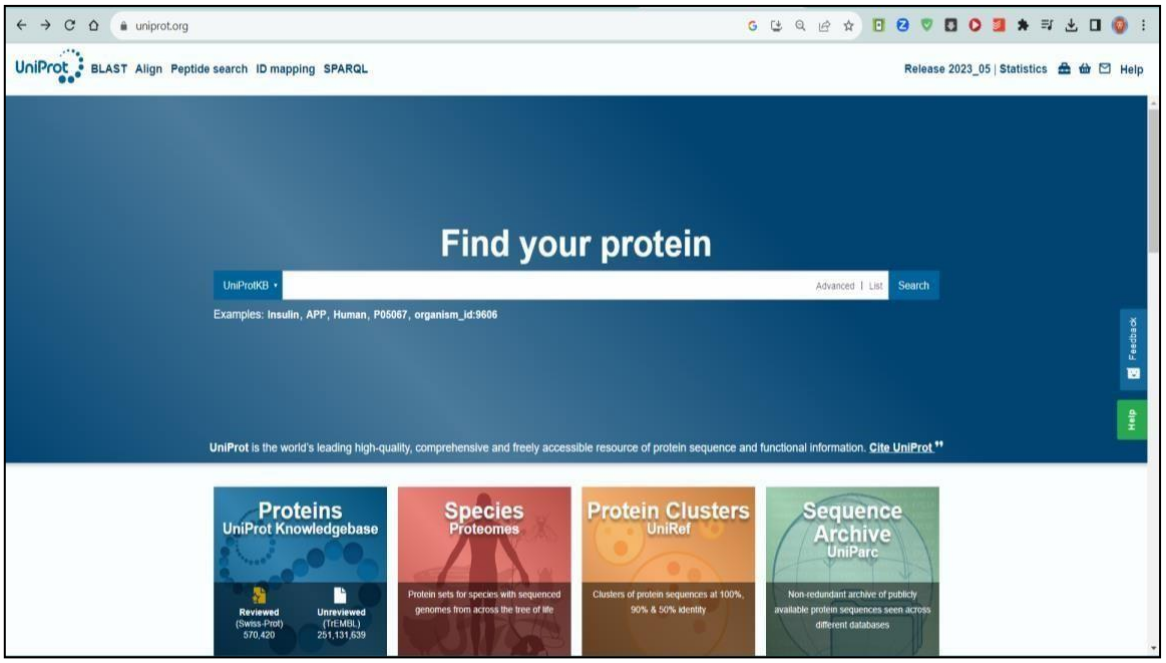


Figure 1: Homepage of the UniProt Database

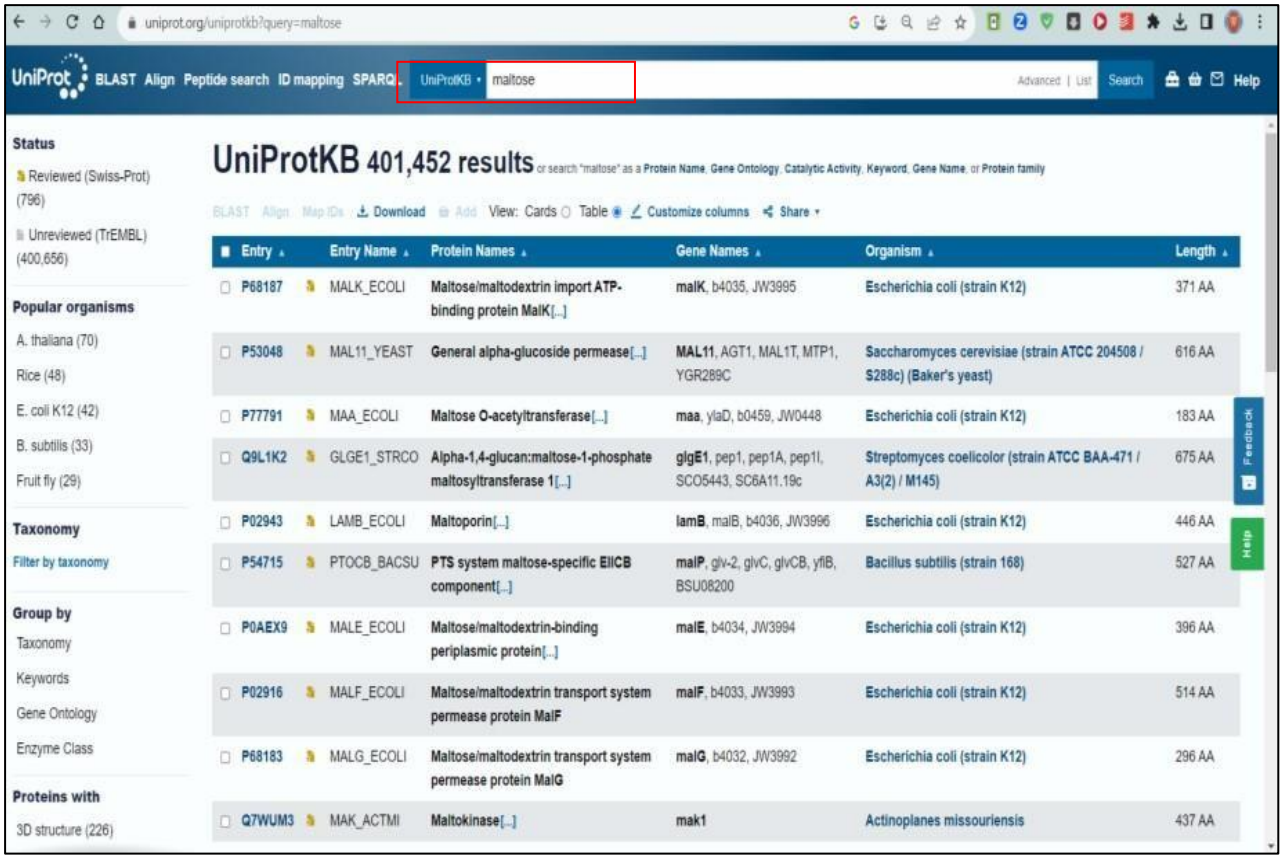


Figure 2: Searching for query maltose protein.

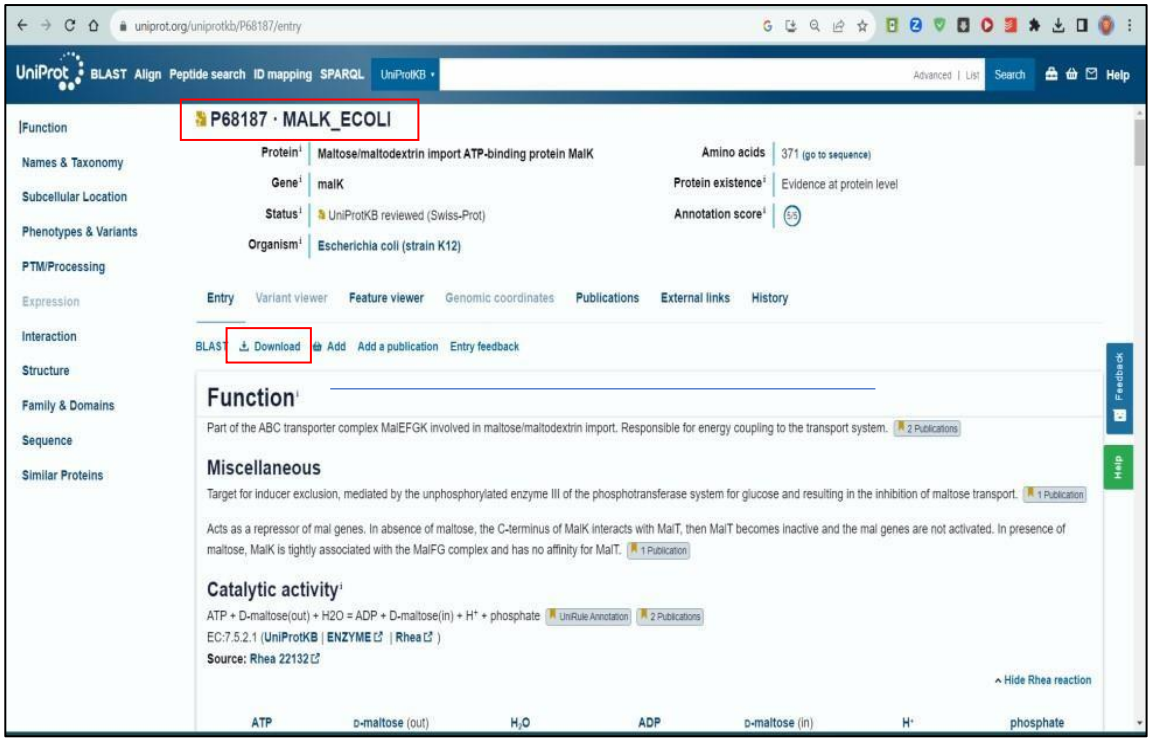


Figure 3: 'Download' option for retrieving the FASTA sequence of the protein

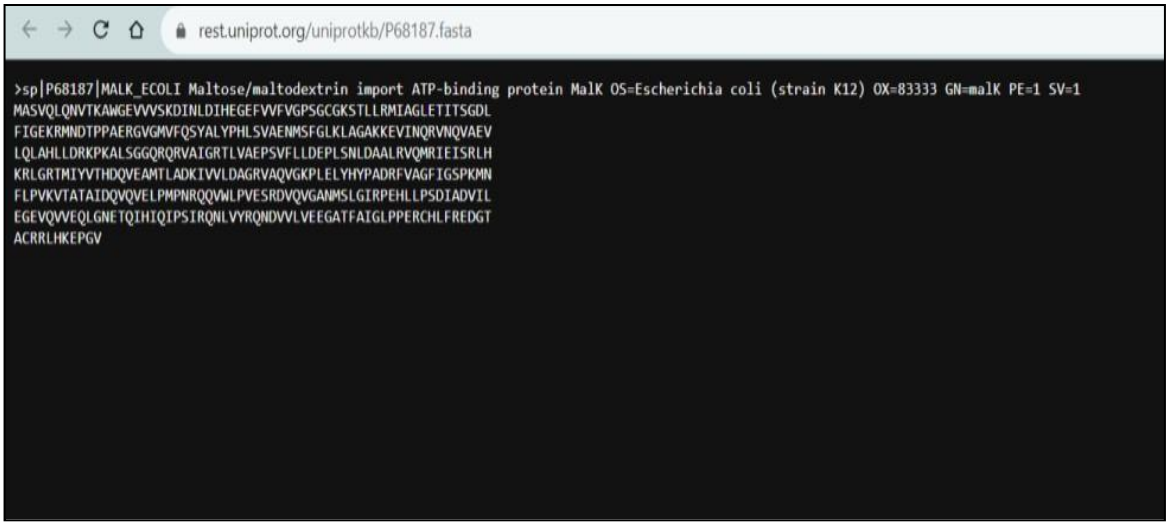


Figure 4: FASTA sequence of maltose protein.

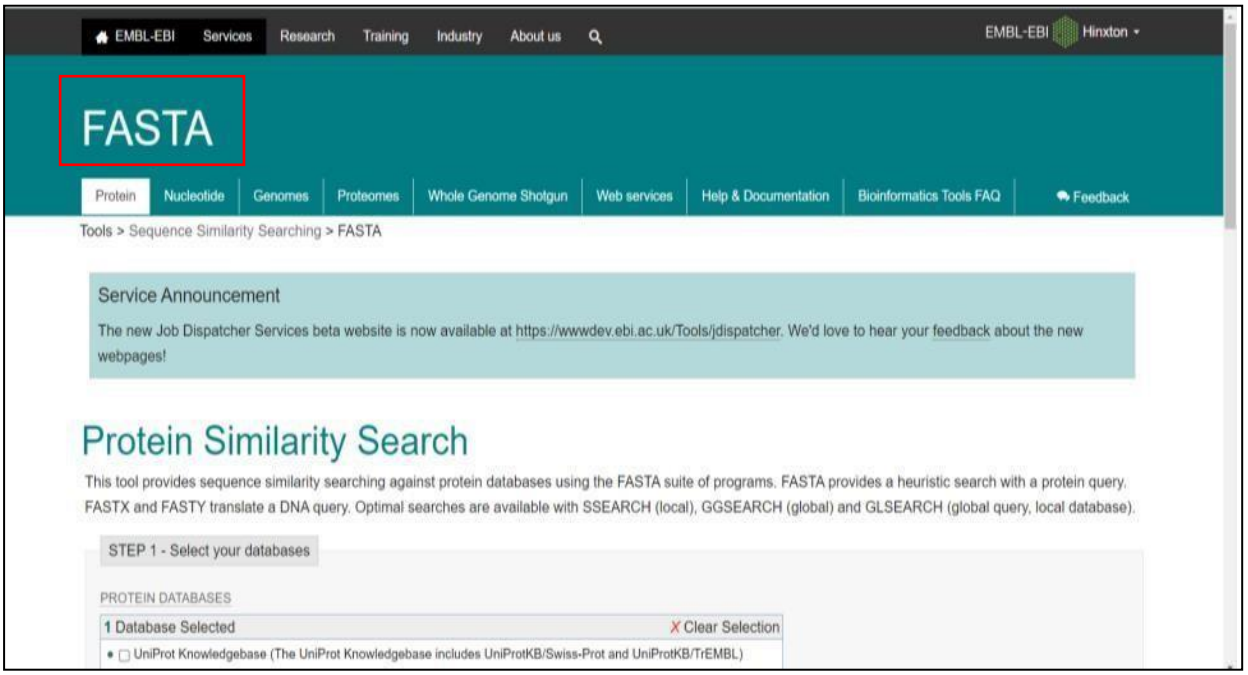


Figure 5: Homepage of FASTA tool.

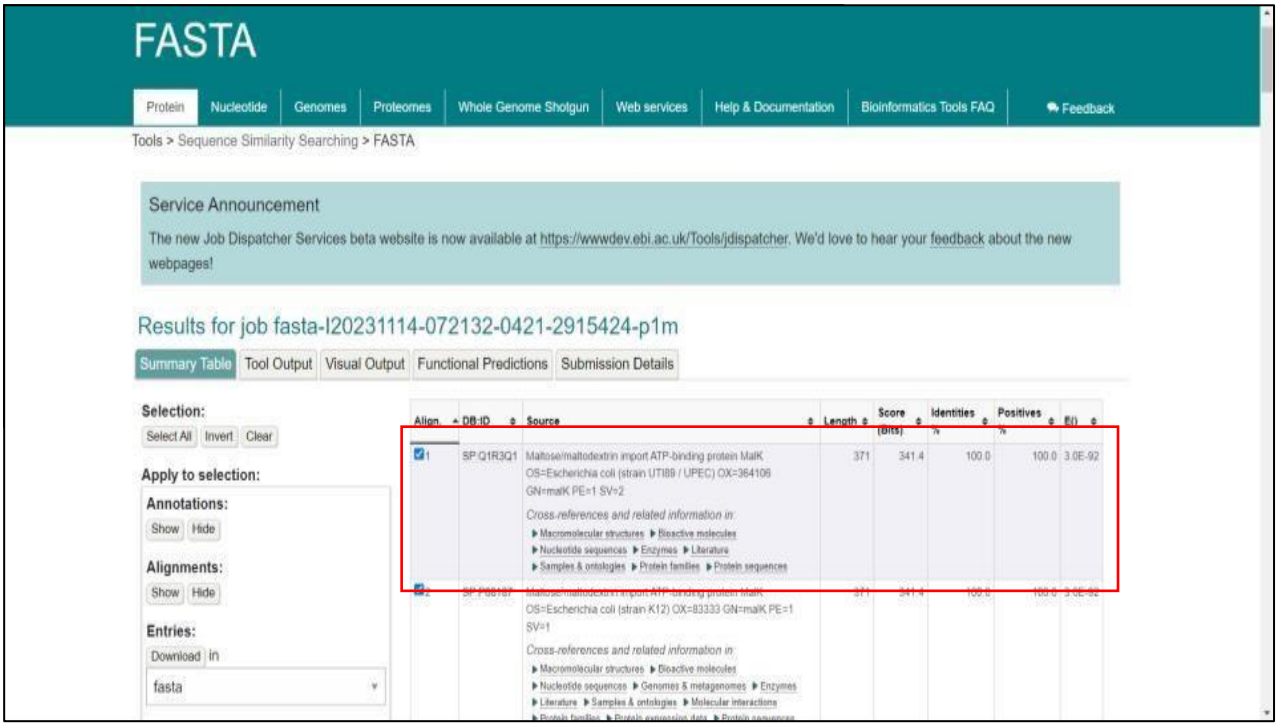


Figure 6: Searching sequence protein in FASTA tool.

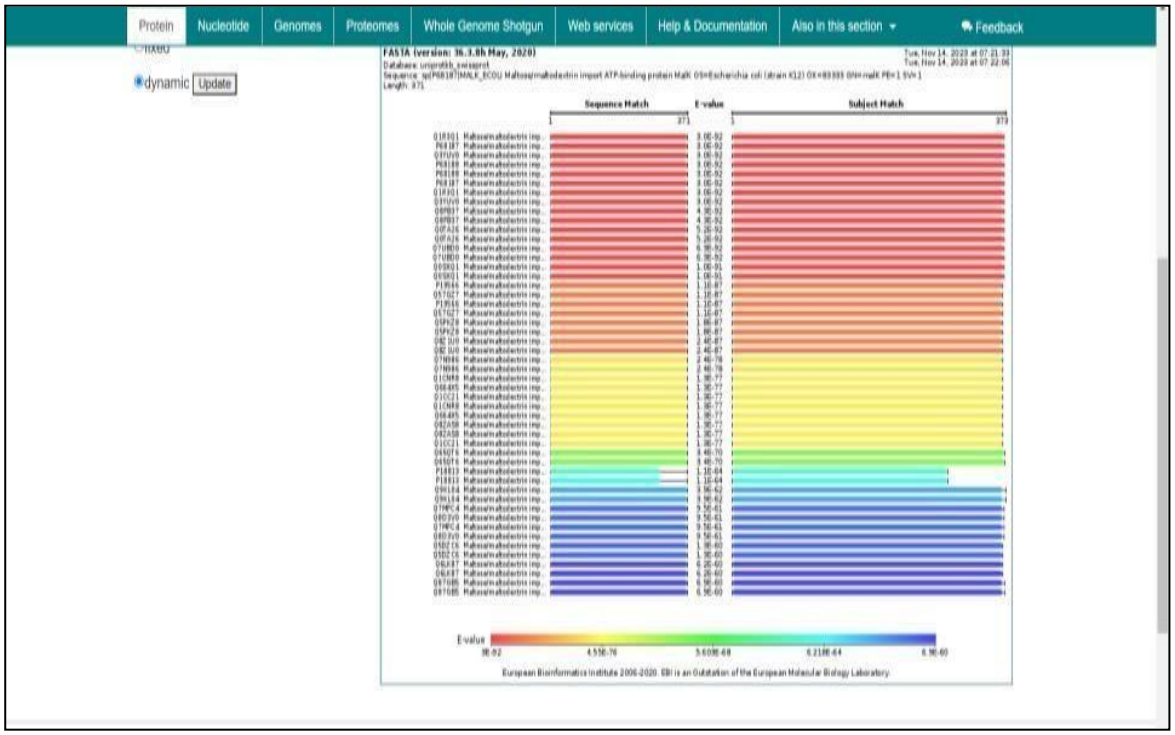


Figure 7: Visual output of maltose protein sequence.

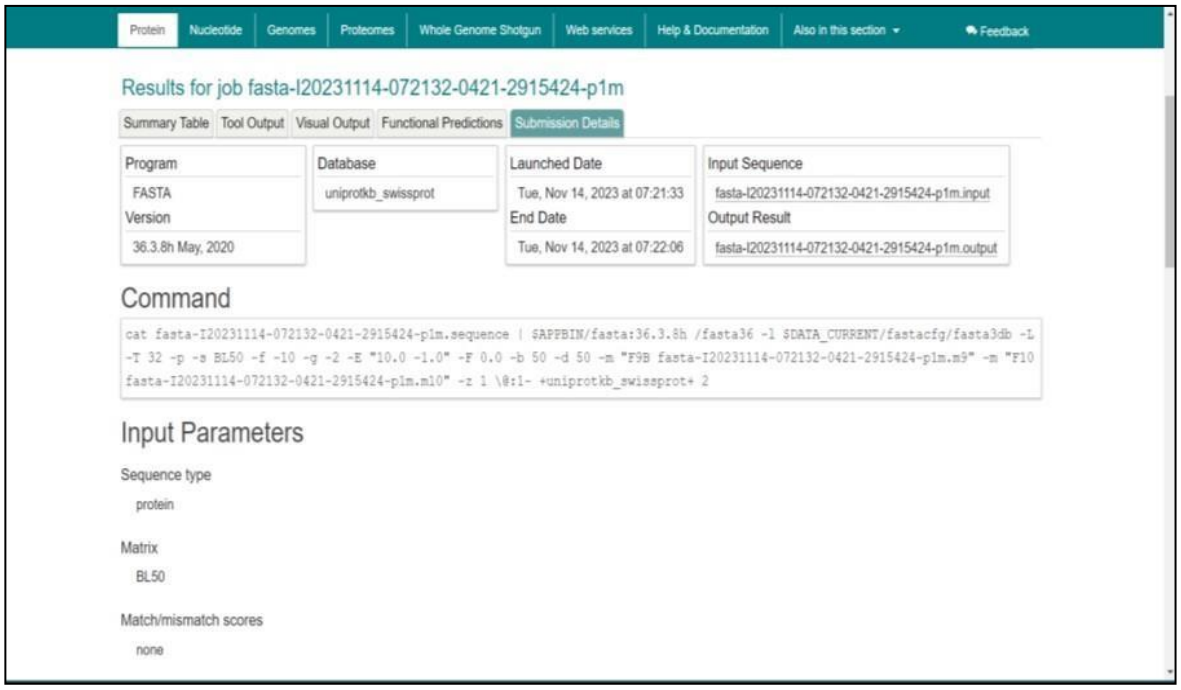


Figure 8: Submission details of maltose protein on FASTA tool.

RESULTS:

The EBI – FASTA tool was used to explore the sequences similar to the sequence of maltose (UniProt ID: P02768). The query sequence is found 100% identities & 100% positives to maltose sequence entries found in two organisms, viz., *Escherichia coli* and *Shigella sonnei*, with E Value of 5.2e-98 and sequence length of 371.

CONCLUSION:

FASTA is a versatile bioinformatics tool primarily used for storing, searching and comparing biological sequence data. It's commonly employed for tasks like sequence alignment, similarity searches and database comparisons. Sequence similarity was searched and studiedfor the Query 'Maltose' (UniProt ID: P68187) using the FASTA program.

REFERENCES:

1. Kryukov K, Ueda MT, Nakagawa S, Imanishi T (July 2020). "Sequence Compression Benchmark (SCB) database—A comprehensive evaluation of reference-free compressors for FASTA-formatted sequences". GigaScience. 9 (7): giaa072. https://doi.org/10.1093/gigascience/giaa072
2. Andrew Lloyd, Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (Methods of Biochemical Analysis, 43), Briefings in Bioinformatics, Volume 2, Issue 4, December 2001, Pages 407–408, https://doi.org/10.1093/bib/2.4.407
3. Pratas D, Hosseini M, Pinho A (2017). "Cryfa: a tool to compact and encrypt FASTA files". 11th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB). Advances in Intelligent Systems and Computing. Vol. 616. Springer. Pp. 305–312. Doi:10.1007/978-3-319-60816-7_37. https://link.springer.com/book/10.1007/978-3-319-60816-7

DATE: 01/11/2023

WEBLEM 6(C)
PROTEIN- SPECIFIC ITERATED BLAST (PSI BLAST)
(URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi)

AIM:

To explore the PSI BLAST tool to search putative homologs for query "Leucine" (UniProt ID: Q8IX15).

INTRODUCTION:

PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) derives a position-specific scoring matrix (PSSM) or profile from the multiple sequence alignment of sequences detected above a given score threshold using protein–protein BLAST. This PSSM is used to further search the database for new matches and is updated for subsequent iterations with these newly detected sequences. Thus, PSI-BLAST provides a means of detecting distant relationships between proteins. BLAST (Basic Local Alignment Search Tool) is a sequence similarity search method, in which a query protein or nucleotide sequence is compared to nucleotide or protein sequences in a target database to identify regions of local alignment and report those alignments that score a given score threshold. Position-Specific Iterative (PSI)-BLAST is a protein sequence profile search method that builds off the alignments generated by a run of the BLASTp program. The first iteration of a PSI-BLAST search is identical to a run of BLASTp program. It then generates a multiple alignment of the highest scoring pairs of the BLASTp run above a certain preset score or *e*-value threshold and calculates a profile or a position-specific score matrix (PSSM) from the multiple alignment.

The PSSM captures the conservation pattern in alignment and stores it as a matrix of scores for each position in the alignment-highly conserved positions receive high scores and weakly conserved positions receive scores near zero. This profile is used in place of the original substitution matrix for a further search of the database to detect sequences that match the conservation pattern specified by the PSSM. The newly detected sequences from this second round of the search, which are above the specified score (*e*-value) threshold is again added to alignment the profile is refined for another round of searching. This process is iteratively continued until desired or until convergence, i.e., the state where no new sequences are detected above the defined threshold. The iterative profile generation process makes PSI-BLAST far more capable of detecting distant sequence similarities than a single query alone in BLASTp, because it combines the underlying conservation information from a range of related sequence into a single score matrix. In the evolution, three-dimensional (3D) structures of proteins may be conserved even after considerable erosion of their sequence similarity. PSI-BLAST has been demonstrated to be useful in detecting such relationships via sequence searches, which were previously only detected through direct comparison of the 3D structures. Here, we discuss practical aspects of using PSI-BLST and provide a tutorial on how to uncover distant relationships between proteins and use them to reach biological meaningful conclusions.

Significance:

1. PSI-BLAST is most conveniently used on the internet with the help of the graphical user interface provided by the PSI-BLAST search page on National Centre for Biotechnology Information (NCBI).
2. The PSI-BLAST page may be customized by the user in terms of automated or semiautomated or "two-page formatting" and other parameters modified as desired. This page can then be saved as permanent internet bookmark for repeated use on futureoccasions.
3. As a rule of the thumb, beginners are advised to use the profile-inclusion threshold of expect (*e*)-value = 0.005 for their analysis. However, a user familiar with globular domains and compositional bias may use the inclusion threshold of 0.01 for inclusion in the profile, if a sequence does not have any major compositionally biased segments.
4. A pair of protein sequences can either be homologous (sharing a common evolutionary ancestor) or nonhomologous (evolutionarily unrelated).
   a. It should be noted that PSI-BLAST does not offer a direct binary decision on whether two sequences are related or not. However, the *e*-value obtained for a PSI-BLAST alignment can be used as a guide for this purpose.
5. As a heuristic it may be assumed that any compositionally unbiased query, encompassing a globular domain in a protein, giving a hit with *e*-value = <0.01 is likelyto be an indication of a homologous relationship. However, a user must carefully evaluate such alignments case-by-case because there can occasionally be false- positives.
6. A user may set the number of alignments and hits view as at least 1000 if searching the nonredundant (nr) database of NCBI, because of the large number hits obtained due tothe current size of the database. PSI-BLAST may also be downloaded and run as a standalone program for Windows or UNIX-type operating systems.
   a. However, in this case the various parameters need to be specified using the set of command-line flags for the program. An advantage of using the standalone version is the ability to use alignments as queries to generate a starting PSSM or saving and reusing the profile generated by a run of PSI-BLAST.

Leucine:

Leucine (symbol **Leu** or **L**) is essential amino acid that is used in the biosynthesis of proteins. Leucine is an α-amino acid, meaning it contains an α-amino group (which is in the protonated $-NH_3^+$ form under biological conditions), an α-carboxylic acid group (which is in the deprotonated $-COO^-$ form under biological conditions), and a side chain isobutyl group, making it a non-polar aliphatic amino acid. It is essential in humans, meaning the body cannot synthesize it: it must be obtained from the diet. Human dietary sources are foods that contain protein, such as meats, dairy products, soy products, and beans and other legumes. It is encoded by the codons UUA, UUG, CUU, CUC, CUA, and CUG.

Like valine and isoleucine, leucine is a branched-chain amino acid. The primary metabolic end products of leucine metabolism are acetyl-CoA and acetoacetate; consequently, it is one of the two exclusively ketogenic amino acids, with lysine being the other. It is the most important ketogenic amino acid in humans.

L-leucine is the L-enantiomer of leucine. It has a role as a plant metabolite, an Escherichia coli metabolite, a Saccharomyces cerevisiae metabolite, a human metabolite, an algal metabolite

and a mouse metabolite. It is a pyruvate family amino acid, a proteinogenic amino acid, a leucine and a L-alpha-amino acid. It is a conjugate base of a L-leucinium. It is a conjugate acid of a L-leucinate. It is an enantiomer of a D-leucine. It is a tautomer of a L-leucine zwitterion.

METHODOLOGY:

1. Go to the website of BLAST tool.
2. Click protein blast as protein is more conserved than nucleotide.
3. Go on UniProt portal.
4. Search for query 'Leucine'.
5. From shown results select UniProt ID: 'Q8IX15' entry.
6. Download the sequence in FASTA (Canonical) format.
7. Copy the sequence and paste under BLASTp suite.
8. Select Protein Data Bank (PDB) database under standard and program algorithm parameter as psi-blast with threshold 0.001.
9. Click BLAST to run the query.
10. Click Run to observe 2nd iterated and continue till 5 iterations.

OBSERVATIONS:

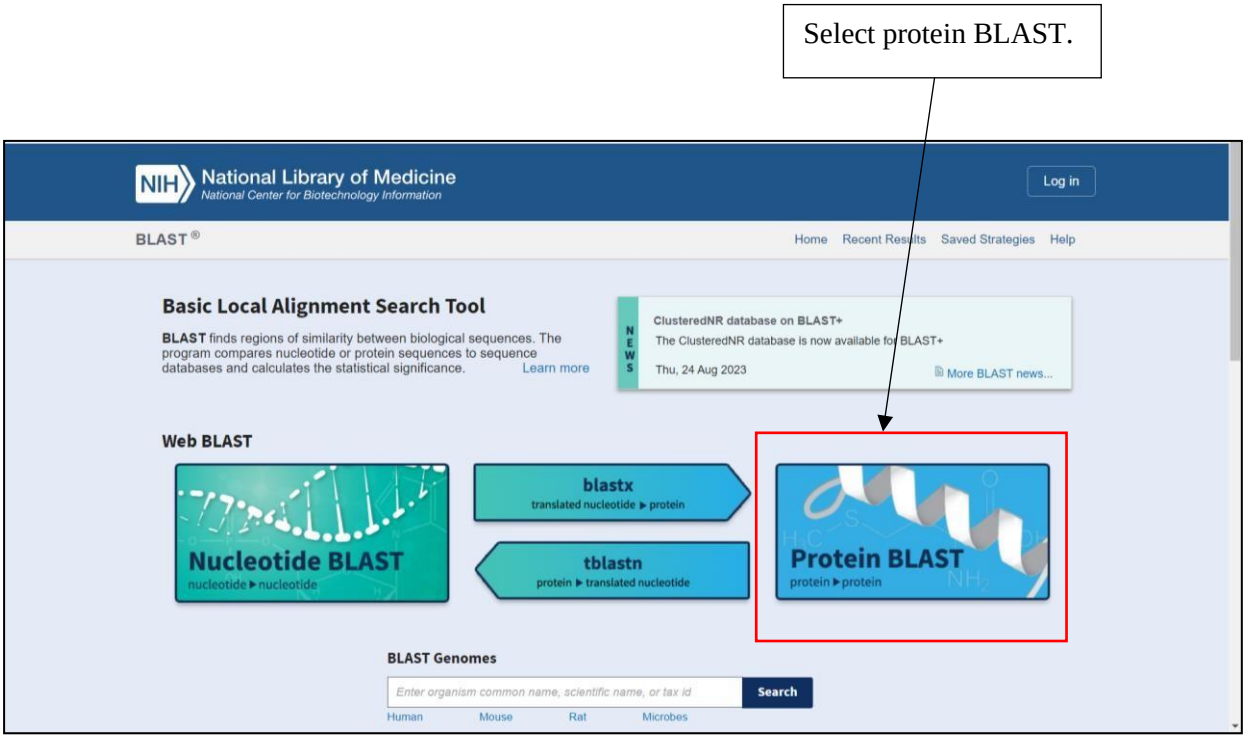


Figure 1: Homepage of BLAST

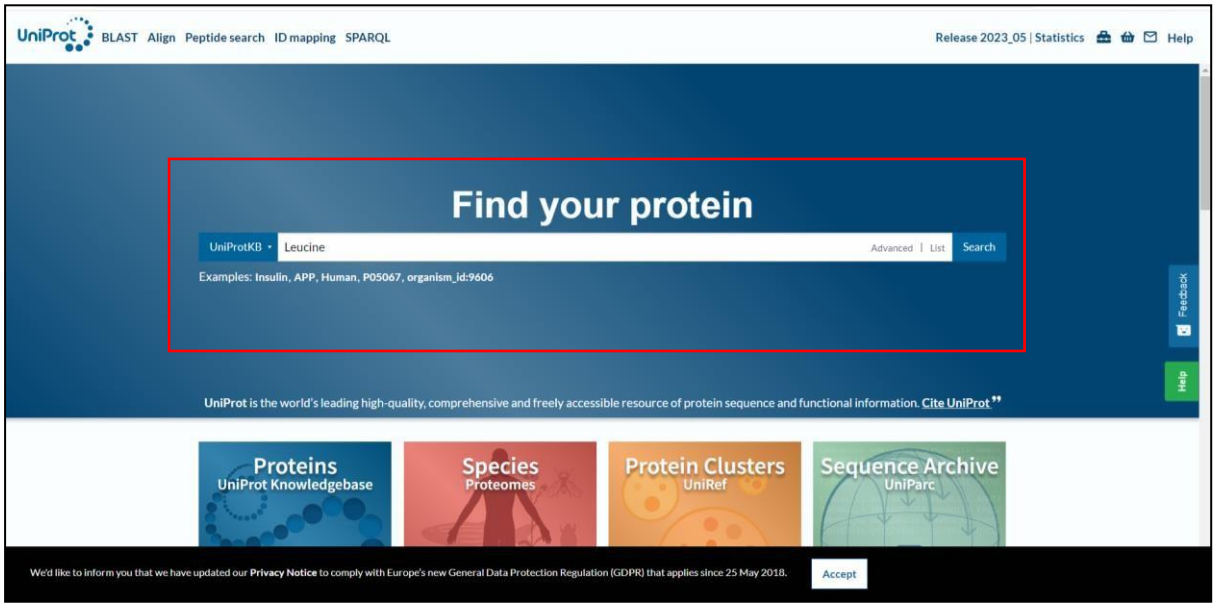


Figure 2: Query search in UniProt portal

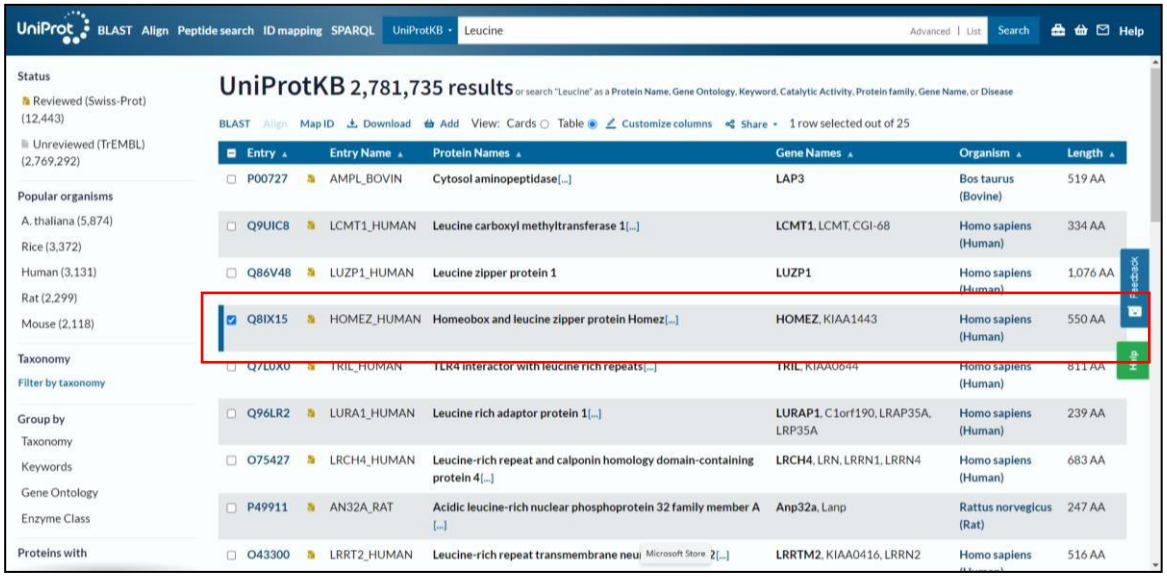


Figure 2a: Select desired organism

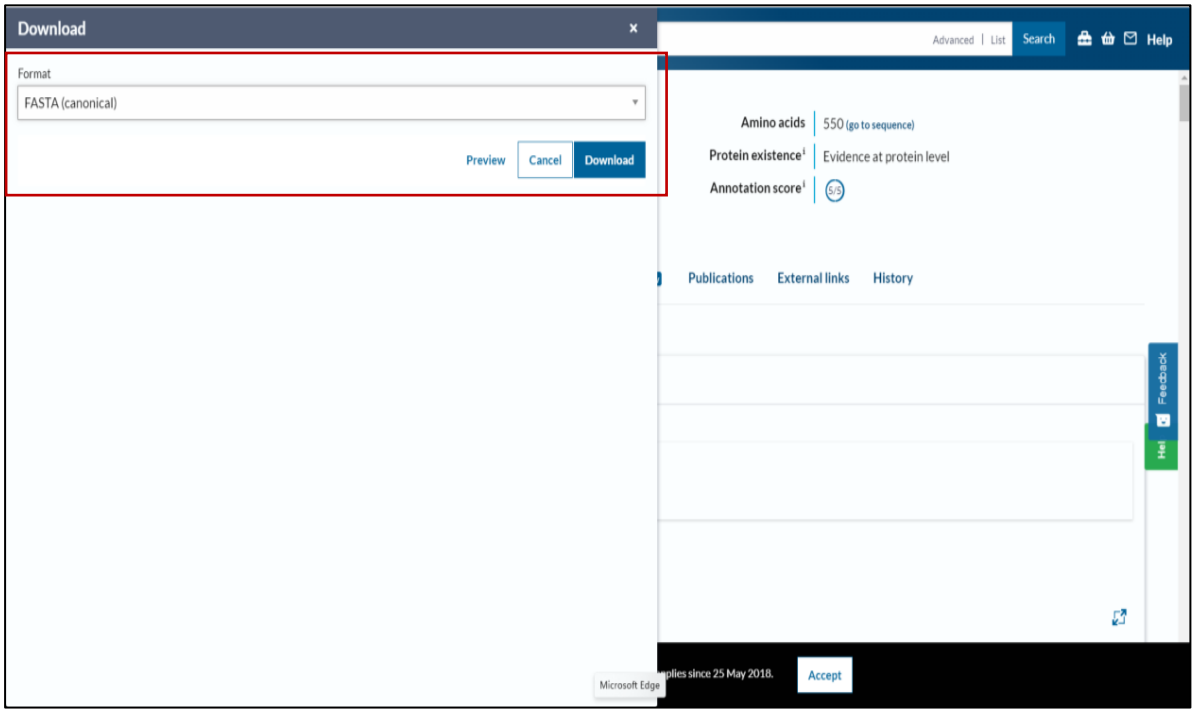


Figure 2b: Download sequence in FASTA (Canonical) format

```
>sp|Q8IX15|HOMEZ_HUMAN Homeobox and leucine zipper protein Homez OS=Homo sapiens OX=9606 GN=HOMEZ PE=1 SV=2
MVRGWEPPPGLDCAISEGHKSEGTMPPNKEASGLSSSPAGLICLPPISEELQLVWTQAAQ
TSELDSNEHLLKTFSYFPYPSLADIALLCLRYGLQMEKVKTWFMAQRLRCGISWSSEEIE
ETRARVVYRRDQLHFKSLLSFTHHAGRPPEEVPPPPVPAPEQVGIGIGPPTLSKPTQTKG
LKVEPEEPSQMPPLPQSHQKLKESLMTPGSGAFPYQSDFWQHLQSSGLSKEQAGRGPNQS
HGIGTASWNHSTTVPQPQARDKPPPIALIASSCKEESASSVTPSSSSTSSSFQVLANGAT
AASKPLQPLGCVPQSVSPSEQALPPHLEPAWPQGLRHNSVPGRVGPTEYLSPDMQRQRKT
KRKTKEQLAILKSFFLQCQWARREDYQKLEQITGLPRPEIIQWFGDTRYALKHGQLKWFR
DNAVPGAPSFQDPAIPTPPPSTRSLNERAETPPLPIPPPPPDIQPLERYWAAHQQLRETD
IPQLSQASRLSTQQVLDWFDSRLPQPAEVVVCLDEEEEEEEEELPEDDEEEEEEEEEDDD
DDDDDVIIQD
```

Figure 2c: Copying the sequence

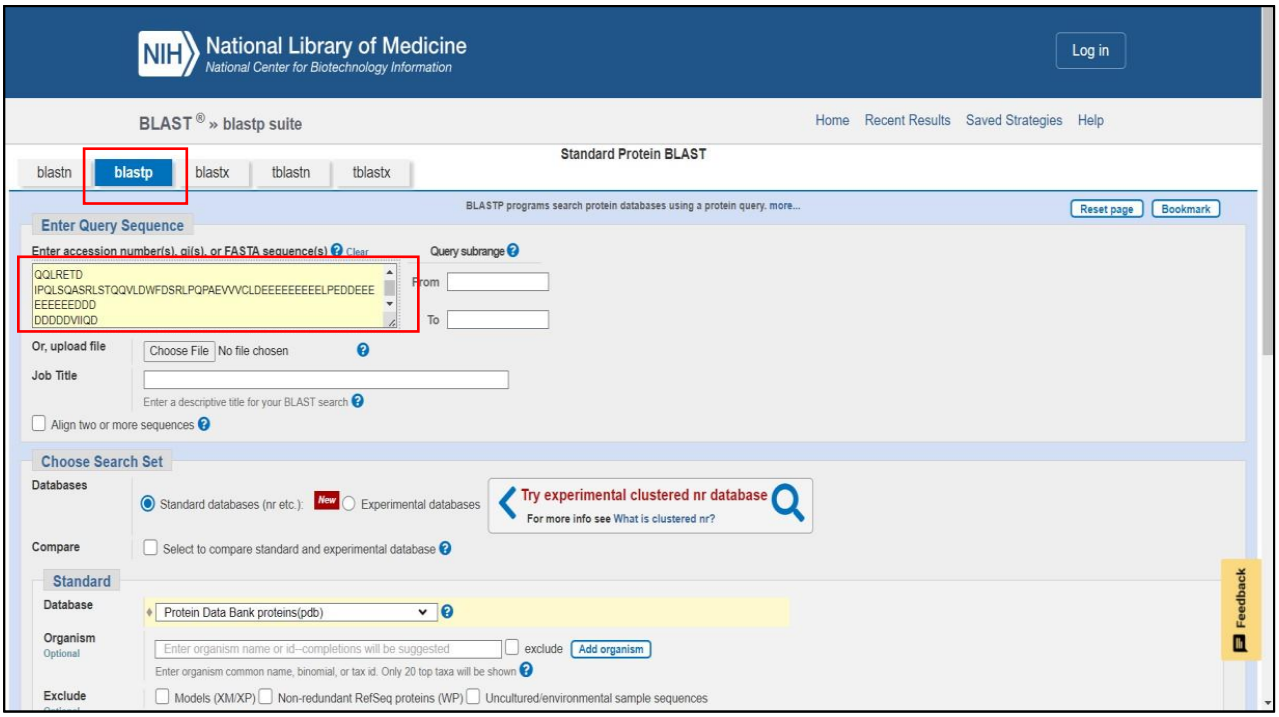


Figure 3: Pasting the sequence in BLASTp format

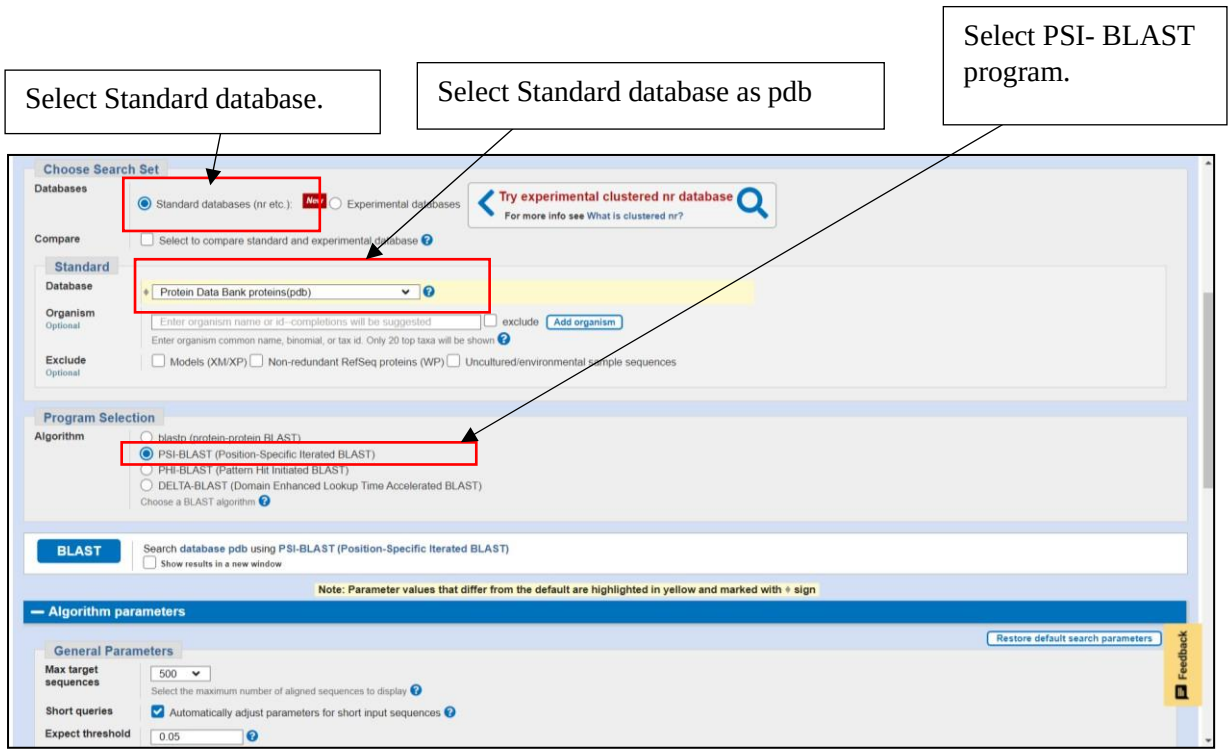


Figure 4: Selecting Standard database as pdb and program selection as PSI- BLAST

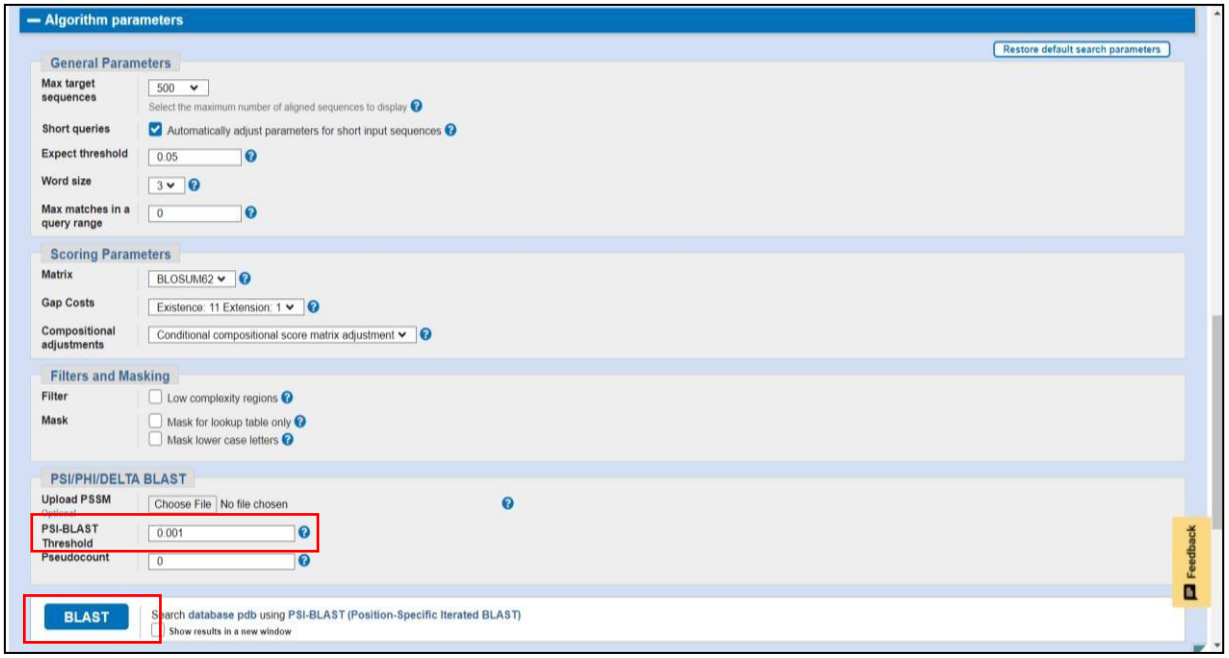


Figure 5: Keeping PSI-BLAST threshold as 0.001 and running PSI - BLAST

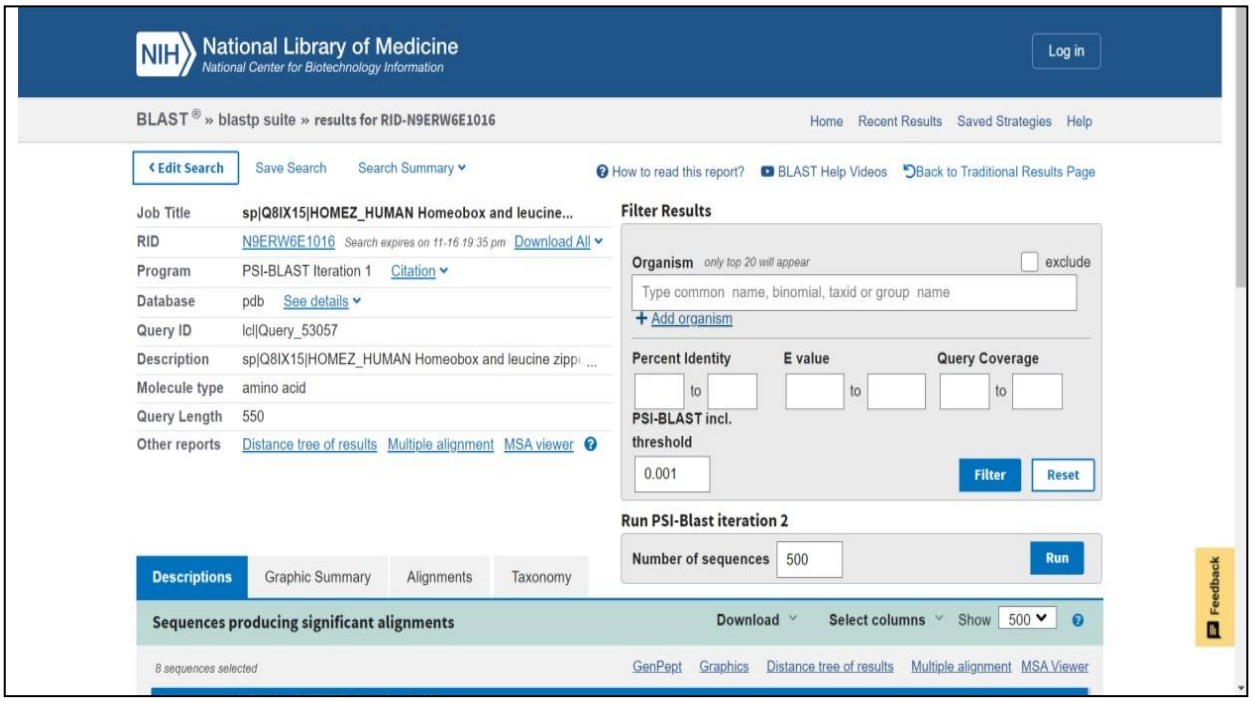


Figure 6: Result shown for UniProt ID: Q8IX15 in BLASTp

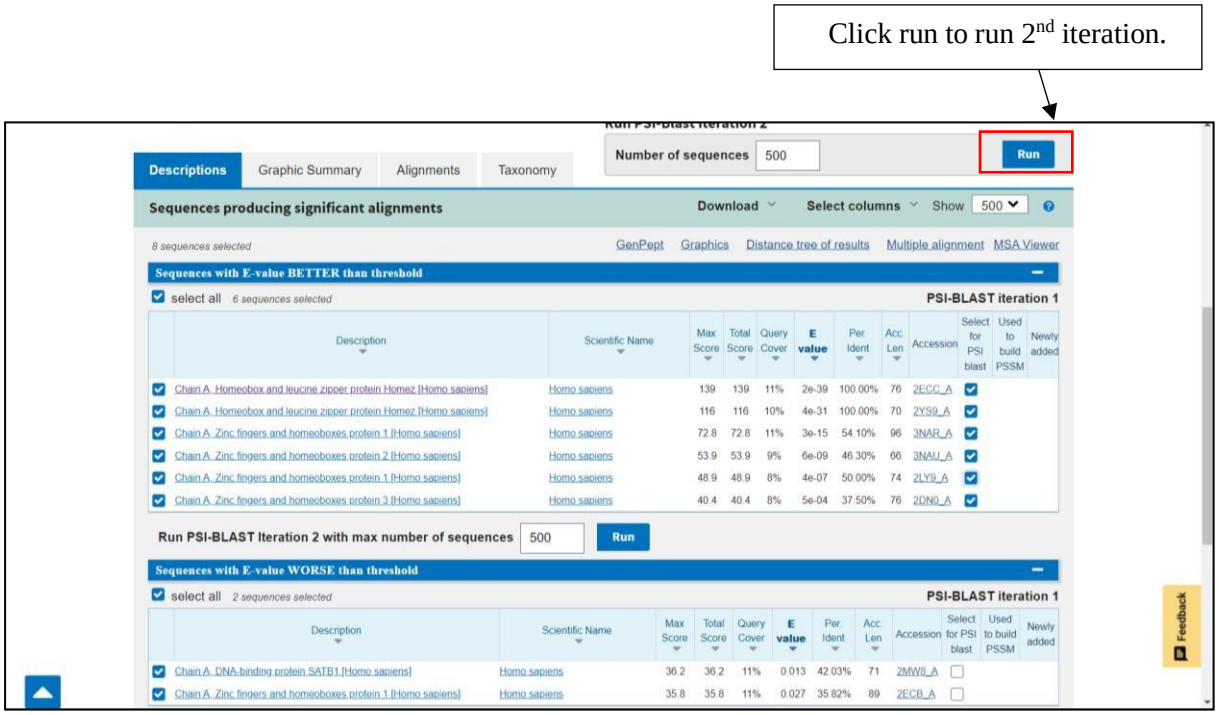


Figure 6a: Result shown for sequence with E- value better and worse than threshold

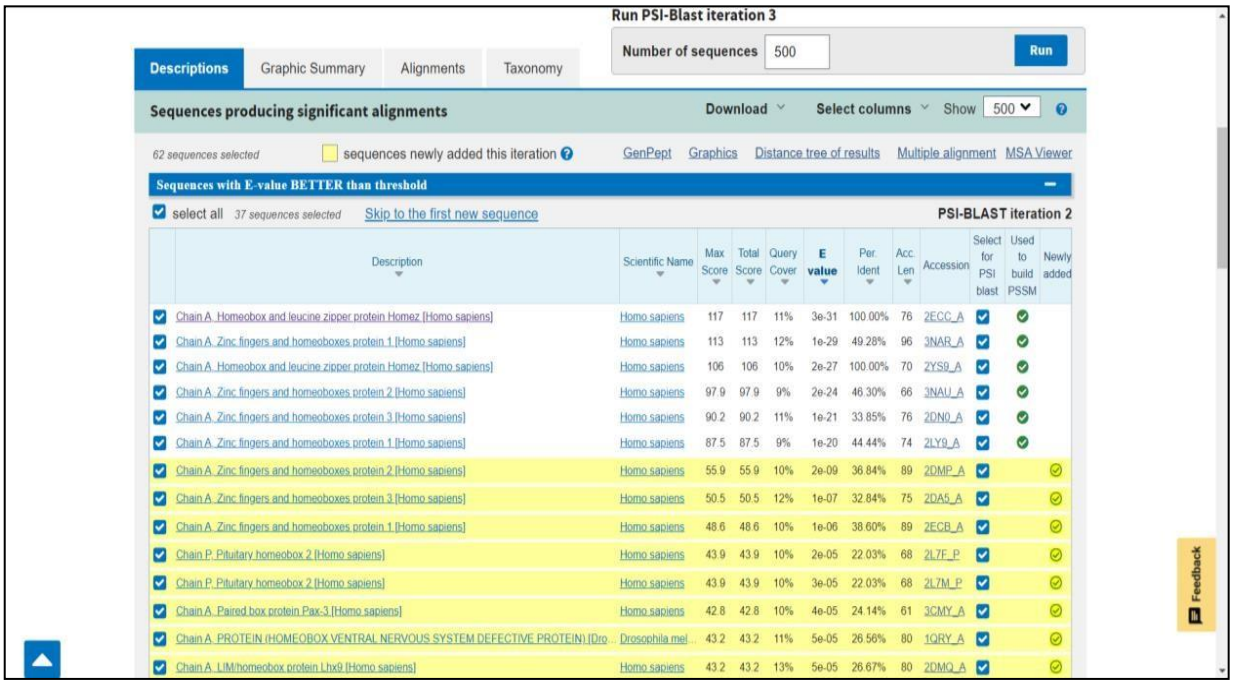


Figure 7: 2nd iterated result of UniProt ID: Q8IX15 organism

RESULTS:

PSI BLAST was explored using query 'Leucine' (Q8IX15) in order to get putative homologs. The first iteration showed 8 new putative sequences and the addition of new sequences was carried till 5th iteration, but then the process if halted as further iteration would drop the result accuracy and the iteration showed that new putative homologs are available for query 'Leucine'.

CONCLUSION:

PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) derives a position-specific scoring matrix (PSSM) or profile from the multiple sequence alignment of sequences detected above a given score threshold using protein–protein BLAST. This PSSM is used to further search the database for new matches and is updated for subsequent iterations with these newly detected sequences. Thus, PSI-BLAST provides a means of detecting distant relationships between proteins. PSI-BLAST (Position specific iterative – BLAST) algorithm program was used to view and explore best iterated results for query 'Leucine' (UniProt ID: Q8IX15).

REFERENCES:

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410.
2. PSI-BLAST. (n.d.). National Institutes of Health. https://www.ncbi.nlm.nih.gov/books/NBK2590/
3. Pruitt KD, Tatusova T, Ostell JM. McEntyre J, Ostell J, editors. The Reference Sequence (RefSeq) Project. National Library of Medicine (US), NCBI; Bethesda, MD: The NCBI Handbook. 2005 Chapter 18.
4. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. (1997 September 01) *Nucleic acids research* 25 (17) :3389-3402
5. Zhang, L., Li, F., Guo, Q., Duan, Y., Wang, W., Zhong, Y., Yang, Y., & Yin, Y. (2020). Leucine Supplementation: A Novel Strategy for Modulating Lipid Metabolism and Energy Homeostasis. *Nutrients*, *12*(5), 1299. https://doi.org/10.3390/nu12051299

DATE: 01/11/2023

WEBLEM 6(D)
PATTERN HIT INITIATED BLAST (PHI-BLAST) TOOL
(URL: https://blast.ncbi.nlm.nih.gov)

AIM:

To perform iterative blast for query 'Flavodoxin' protein (UniProt ID: P53554) by exploring Pattern Hit Initiated BLAST (PHI-BLAST) Tool.

INTRODUCTION:

Pattern Hit Initiated BLAST (PHI-BLAST) Tool, represents a variant of the BLAST algorithm employed for searching a protein database to identify other instances of a specific pattern occurring at least once within the input sequence. It facilitates the alignment and construction of the Position-Specific Scoring Matrix (PSSM) around a motif present in the query sequence. PHI-BLAST was developed by Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, and David J. Lipmann at the National Institutes of Health (NIH).

PHI-BLAST finds application in the analysis of various protein sequences, including CED4-like cell death regulators, HS90-type ATPase domains, archaeal tRNA nucleotidyltransferases, and archaeal proteins. It is utilized to identify protein sequences containing a specific pattern specified by the user and similar to the query sequence.

Compared to other BLAST tools, PHI-BLAST offers advantages such as increased speed and the ability for the user to express a rigid pattern occurrence requirement. This feature aids in reducing the number of hits that solely contain the pattern but lack true homology to the query sequence. However, PHI-BLAST may have a potential disadvantage in that it might be less sensitive than PSI-BLAST for detecting remote homologs. Additionally, the use of a specific pattern may restrict the search scope, potentially causing the omission of homologs lacking the specified pattern.

Flavodoxin:

Flavodoxins are small, soluble, electron-transfer proteins. Flavodoxins contains flavin mononucleotide as prosthetic group. The structure of flavodoxin is characterized by a five-stranded parallel beta sheet, surrounded by five alpha helices. They have been isolated from prokaryotes, cyanobacteria, and some eukaryotic algae. It functions in various metabolic processes, including photosynthesis, nitrogen and fatty acid metabolism. Flavodoxin is also involved in the detoxification of reactive oxygen species. The protein is reduced by flavodoxin reductase and transfers electrons to various redox enzymes. The semiquinone conformation of flavodoxin is stabilized by a hydrogen bond to the N-5 position of flavin, and a common tryptophan residue near the binding site aids in lowering SQ reactivity. The hydroquinone form is forced into a planar conformation, destabilizing it.

METHODOLOGY:

1. Open the homepage of UniProt database and search for the query 'Flavodoxin' protein.
2. Select any one entry from the results e.g., *Bacillus subtilis (strain 168)* **(**UniProt ID: P53554) and download its FASTA sequence in canonical format.
3. Open the homepage of BLAST and click on protein BLAST.
4. Paste the FASTA sequence in 'Enter query sequence' box and in program selection click on PHI-BLAST option.
5. Open the homepage of PROSITE database and search for the query 'Flavodoxin' protein.
6. Enter the FASTA sequence in 'Quick Scan mode of ScanProsite' box and scan it.
7. Copy the decoded pattern and paste it in the pattern in 'Enter a PHI pattern' box on PHI-BLAST portal and set the desired algorithm parameters.
8. Run the PHI-BLAST.
9. After each iteration, the new sequences are added to the results. These new sequences are highlighted using yellow color.
10. Run the PHI-BLAST iteration for 3-5 times, post which it starts generating garbage results, due to the decrease in sensitivity.
11. Interpret the results obtained.

OBSERVATIONS:

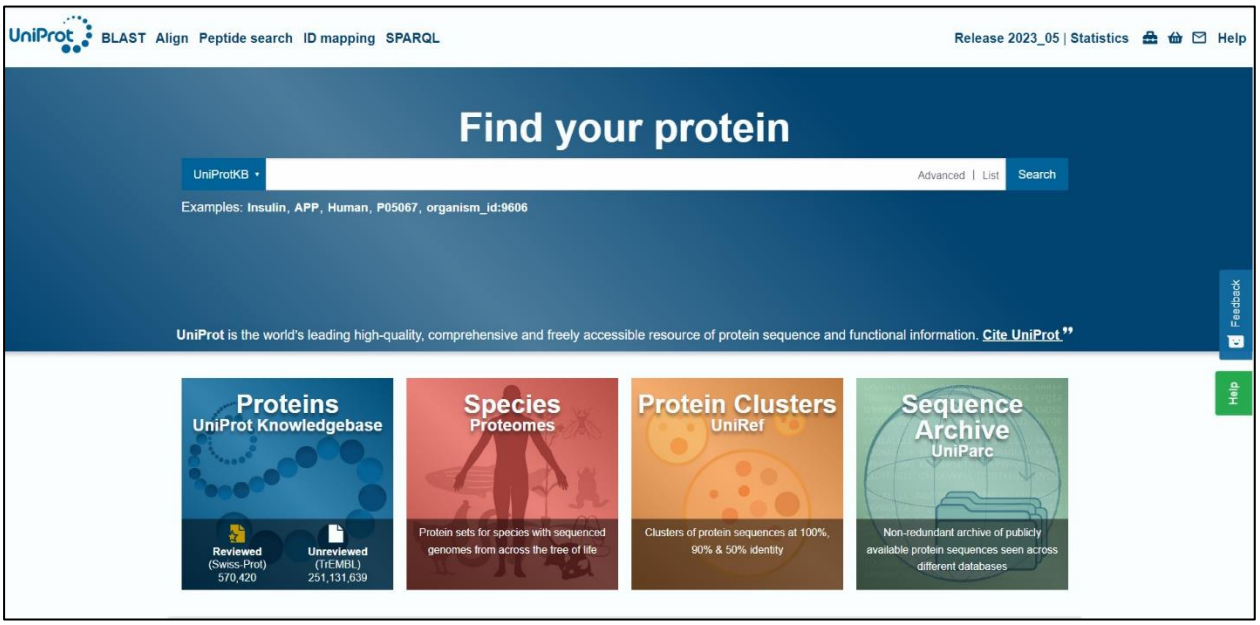


Figure 1: Homepage of the UniProt database

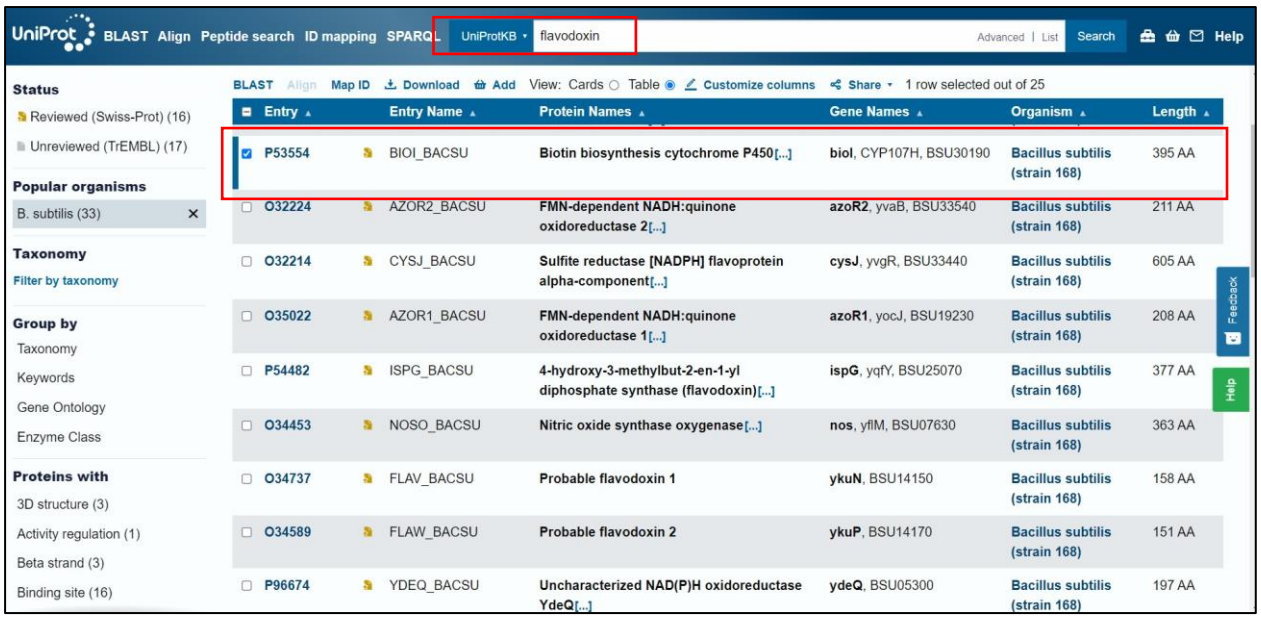


Figure 2: Query search for 'Flavodoxin' protein

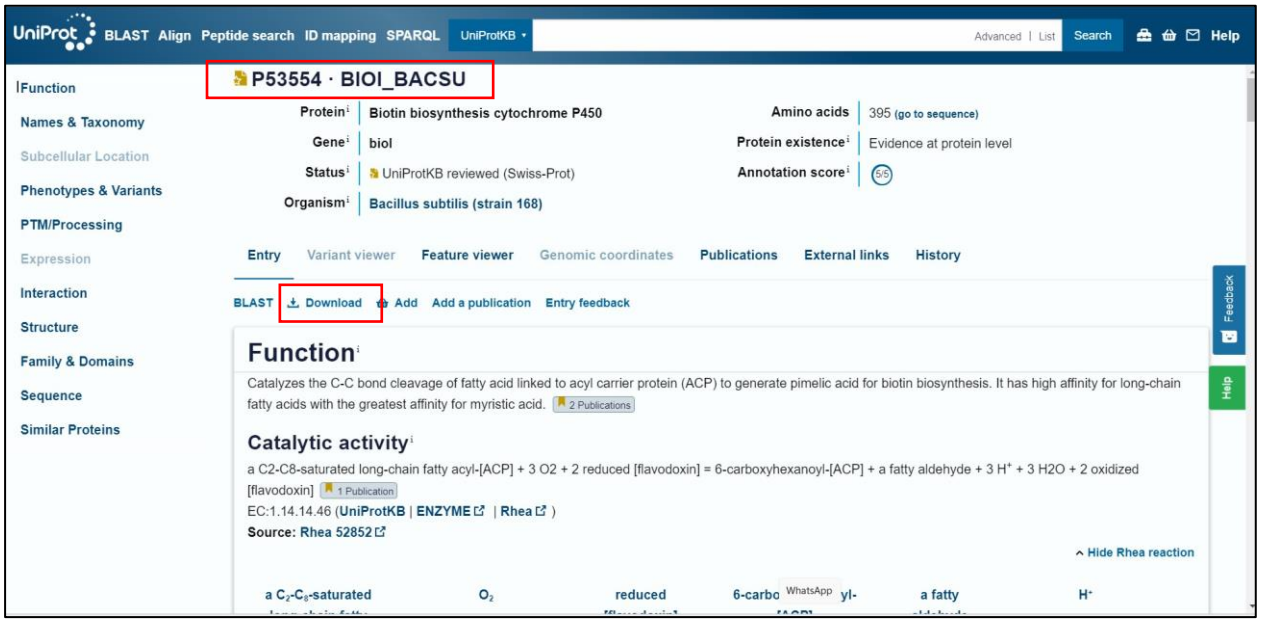


Figure 2a: Downloading the FASTA sequence for selected UniProt ID: P53554

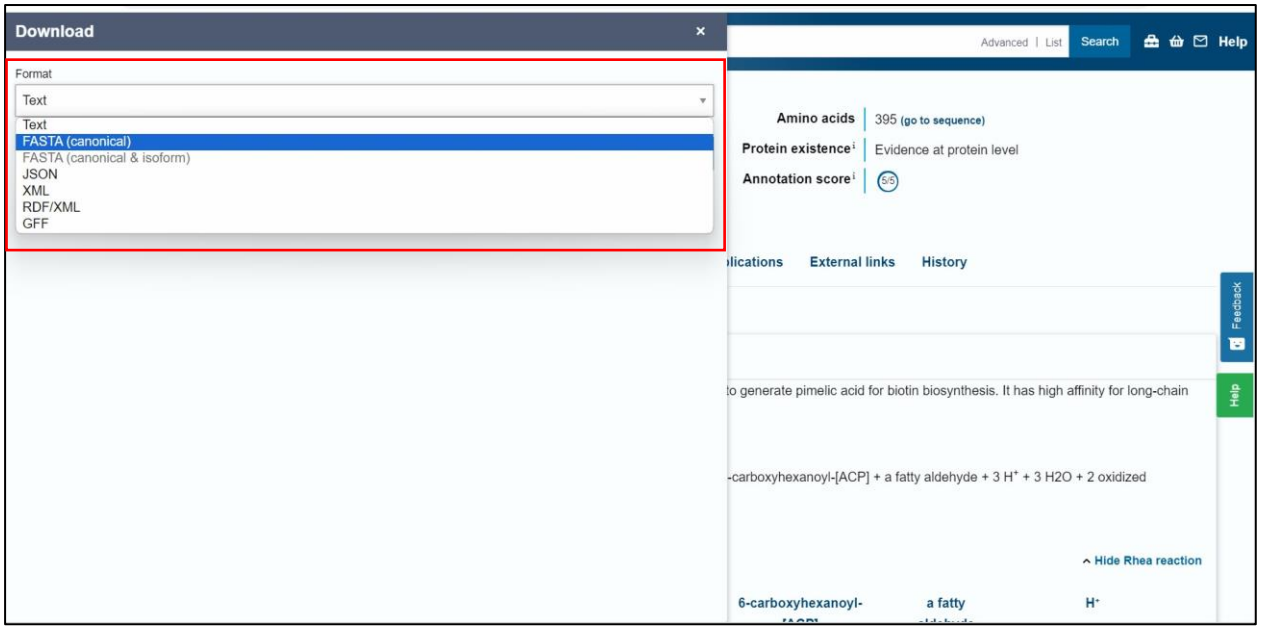


Figure 2b: Downloading the FASTA sequence in canonical format

```
>sp|P53554|BIOI_BACSU Biotin biosynthesis cytochrome P450 OS=Bacillus subtilis (strain 168) OX=224308 GN=bioI PE=1 SV=1
MTIASSTASSEFLKNPYSFYDTLRAVHPIYKGSFLKYPGWYVTGYEETAAILKDARFKVR
TPLPESSTKYQDLSHVQNQMMLFQNQPDHRRLRTLASGAFTPRTTESYQPYIIETVHHLL
DQVQGKKKMEVISDFAFPLASFVIANIIGVPEEDREQLKEWAASLIQTIDFTRSRKALTE
GNIMAVQAMAYFKELIQKRKRHPQQDMISMLLKGREKDKLTEEEAASTCILLAIAGHETT
VNLISNSVLCLLQHPEQLLKLRENPDLIGTAVEECLRYESPTQMTARVASEDIDICGVTI
RQGEQVYLLLGAANRDPSIFTNPDVFDITRSPNPHLSFGHGHHVCLGSSLARLEAQIAIN
TLLQRMPSLNLADFEWRYRPLFGFRALEELPVTFE
```

Figure 2c: View of the downloaded FASTA sequence

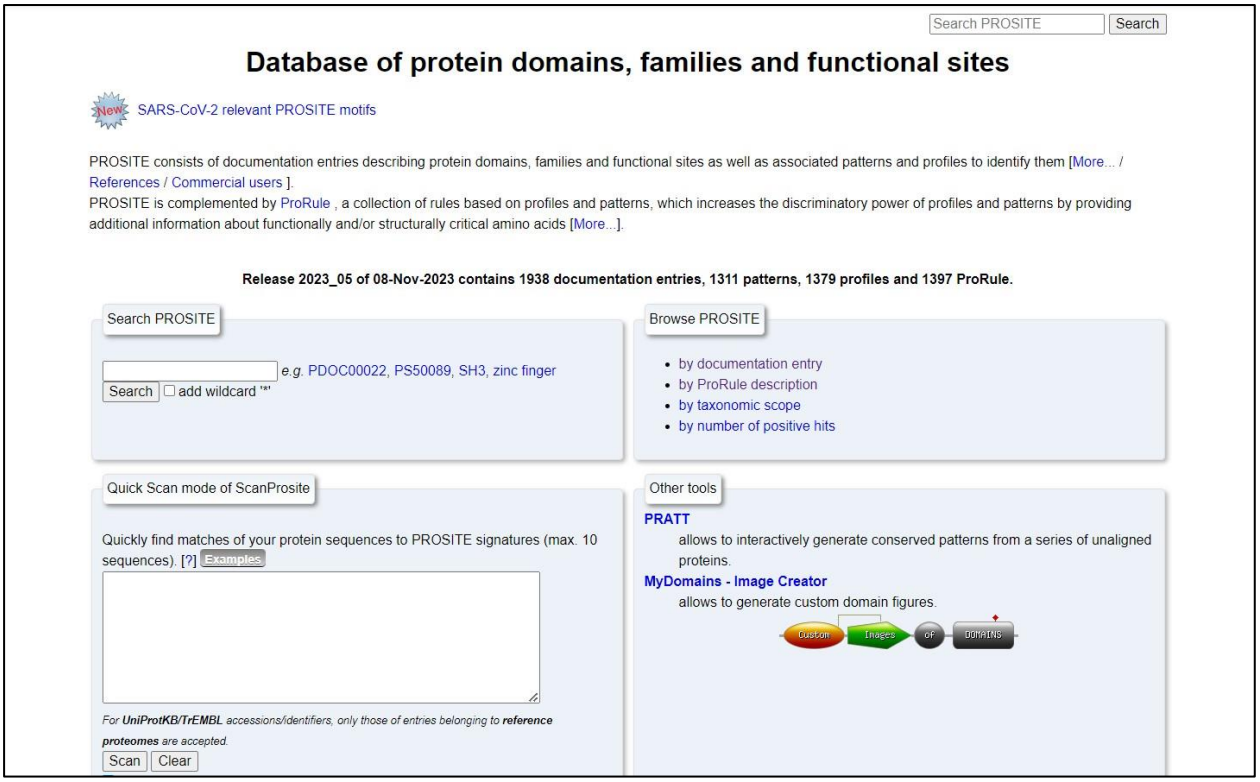


Figure 3: Homepage of PROSITE Database

Figure 3a: Paste the downloaded FASTA sequence for pattern

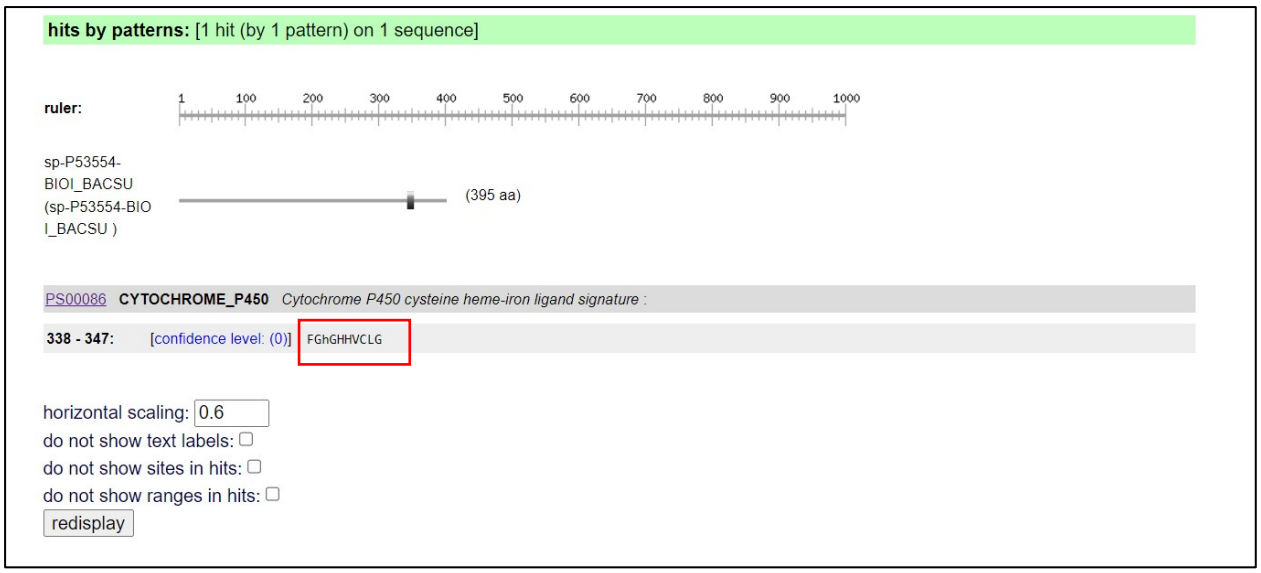


Fig 3b: Results page for the Quick Scan of ScanProSite using the sequence and retrieving the decoded sequence

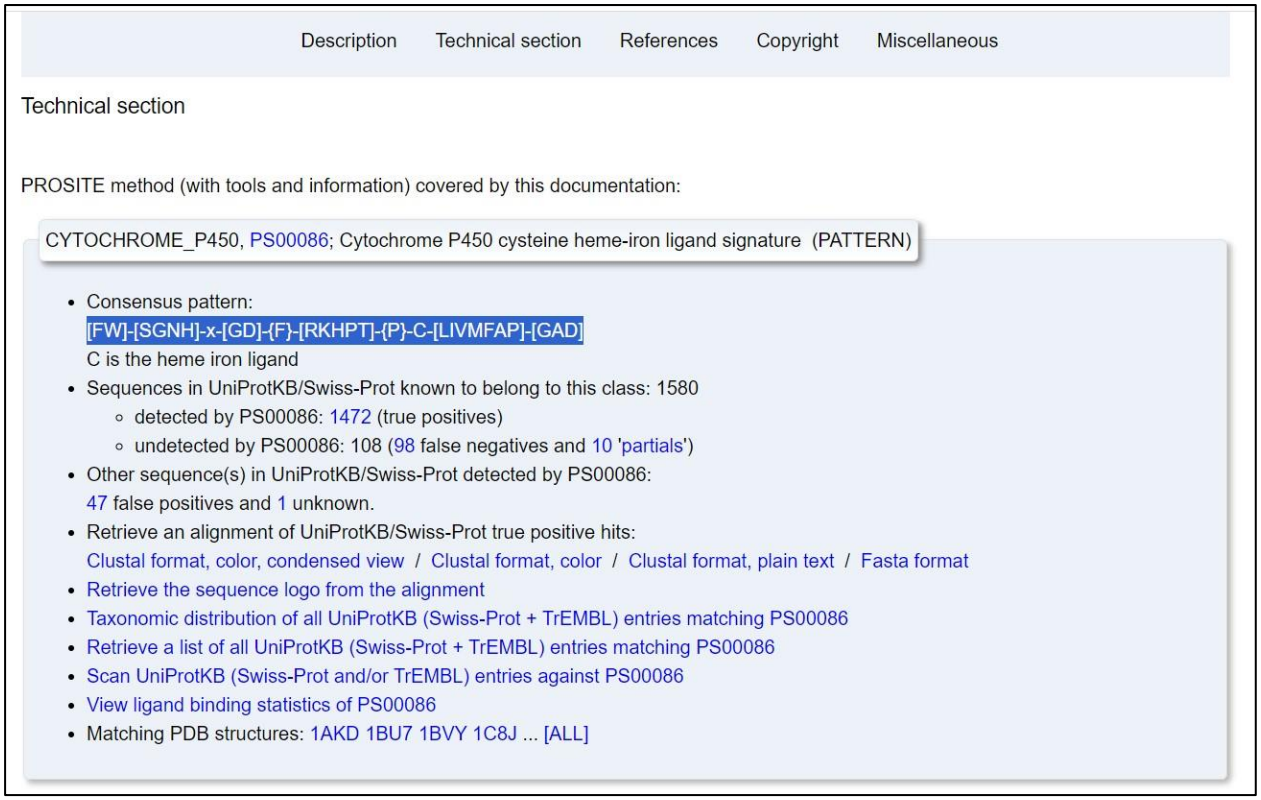


Figure 3c: Consensus pattern for the FASTA sequence

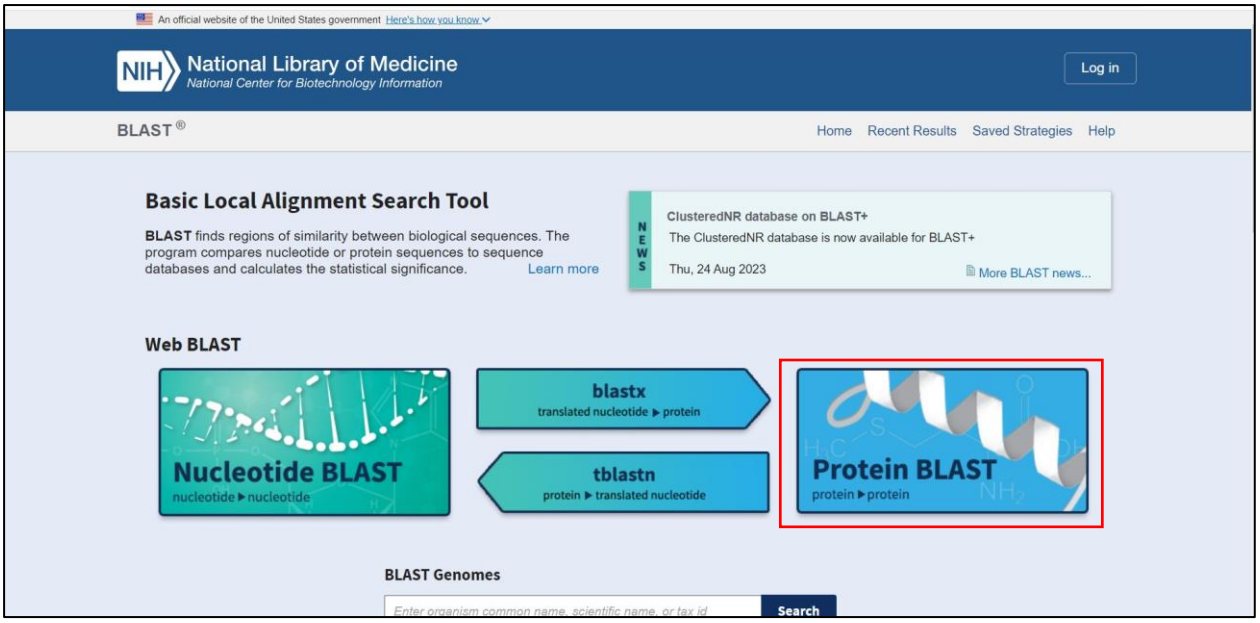


Figure 4: Homepage of Basic Local Alignment Search Tool (BLAST)

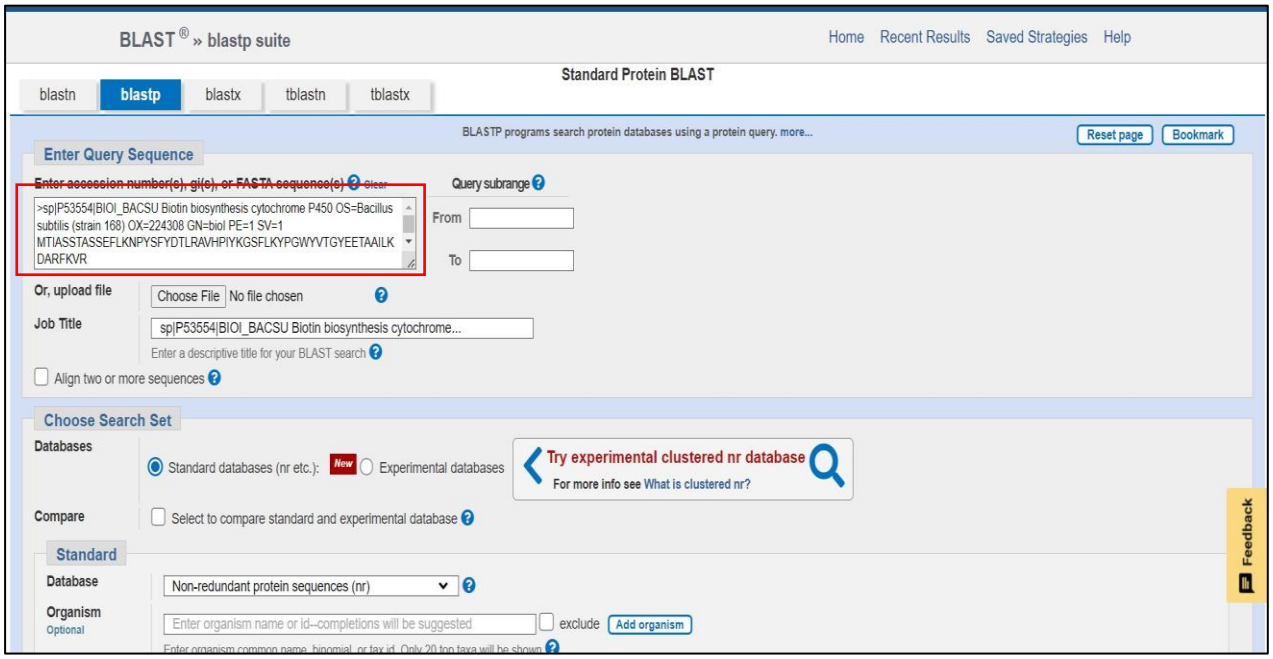


Figure 5: Pasting the FASTA sequence in 'Enter query sequence' box

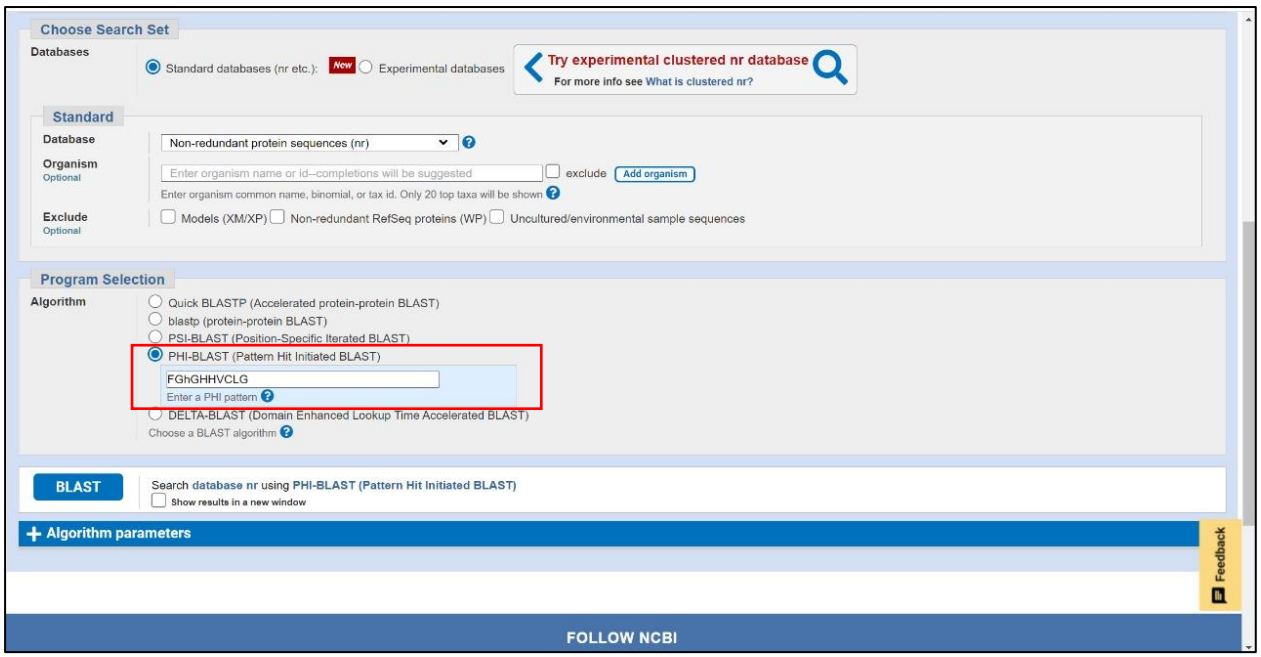


Fig 5a: Paste the decoded pattern from ProSite in 'Enter a PHI pattern' box

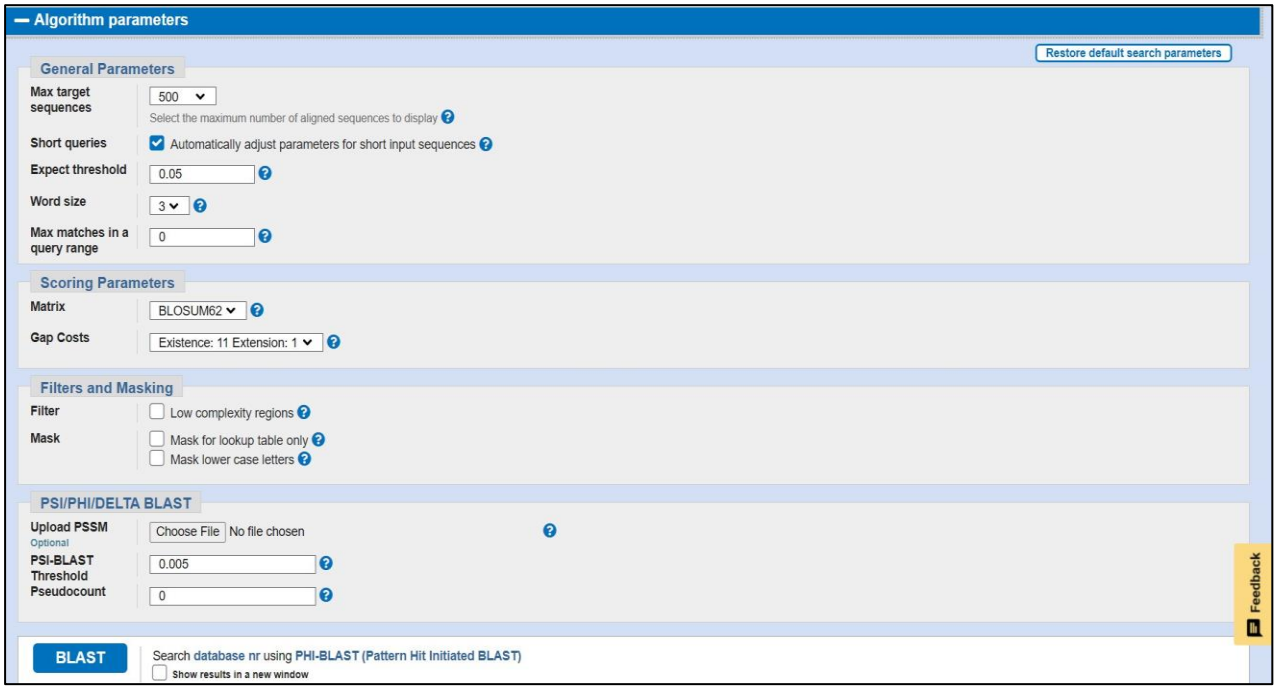


Figure 5b: Setting the parameters for running BLAST Tool

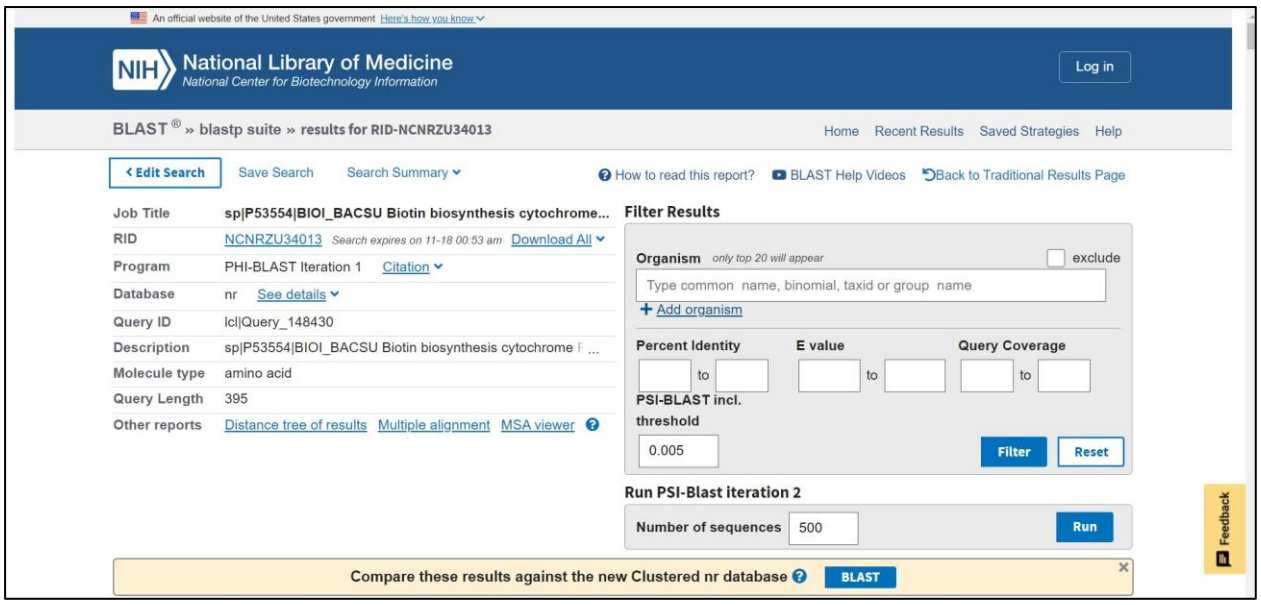
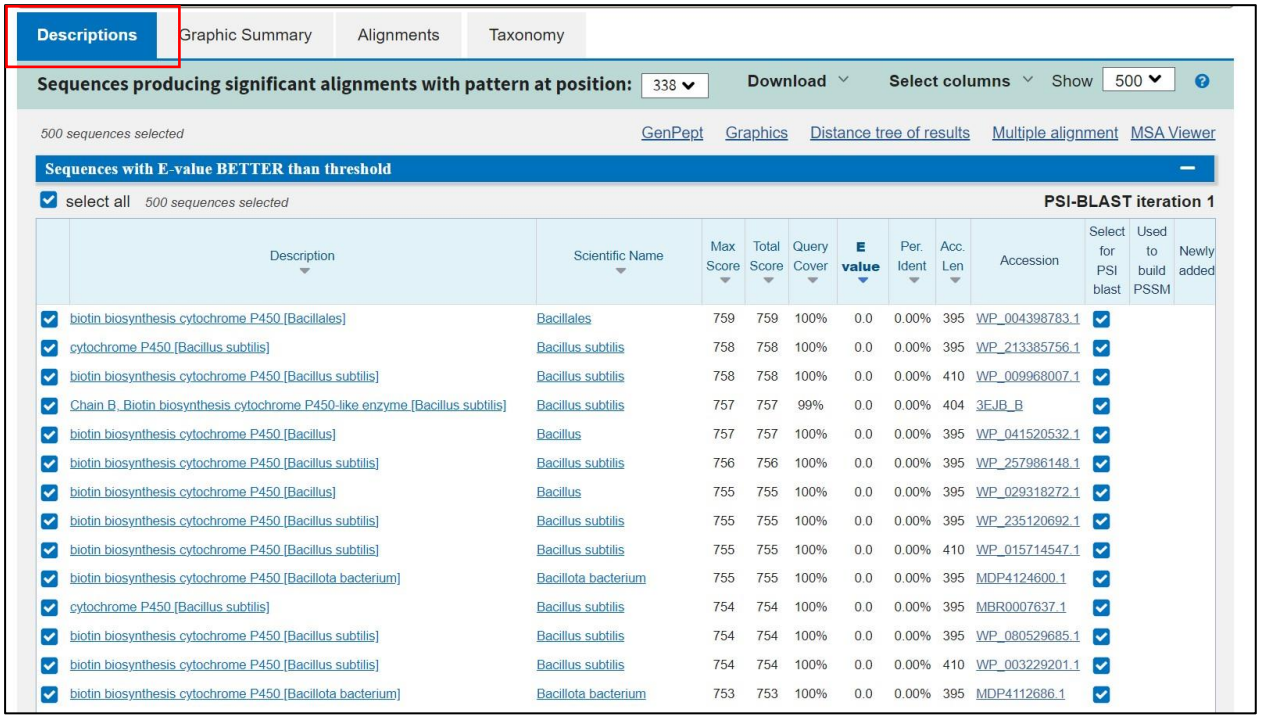


Figure 6: Results obtained after running BLAST tool



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
biotin biosynthesis cytochrome P450 [Bacillales]	Bacillales	759	759	100%	0.0	0.00%	395	WP_004398783.1	✓		
cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	758	758	100%	0.0	0.00%	395	WP_213385756.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	758	758	100%	0.0	0.00%	410	WP_009968007.1	✓		
Chain B, Biotin biosynthesis cytochrome P450-like enzyme [Bacillus subtilis]	Bacillus subtilis	757	757	99%	0.0	0.00%	404	3EJB_B	✓		
biotin biosynthesis cytochrome P450 [Bacillus]	Bacillus	757	757	100%	0.0	0.00%	395	WP_041520532.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	756	756	100%	0.0	0.00%	395	WP_257986148.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus]	Bacillus	755	755	100%	0.0	0.00%	395	WP_029318272.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	755	755	100%	0.0	0.00%	395	WP_235120692.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	755	755	100%	0.0	0.00%	410	WP_015714547.1	✓		
biotin biosynthesis cytochrome P450 [Bacillota bacterium]	Bacillota bacterium	755	755	100%	0.0	0.00%	395	MDP4124600.1	✓		
cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	754	754	100%	0.0	0.00%	395	MBR0007637.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	754	754	100%	0.0	0.00%	395	WP_080529685.1	✓		
biotin biosynthesis cytochrome P450 [Bacillus subtilis]	Bacillus subtilis	754	754	100%	0.0	0.00%	410	WP_003229201.1	✓		
biotin biosynthesis cytochrome P450 [Bacillota bacterium]	Bacillota bacterium	753	753	100%	0.0	0.00%	395	MDP4112686.1	✓		

Figure 7: Result for Description section of query

Figure 8: Result for Graphic Summary section

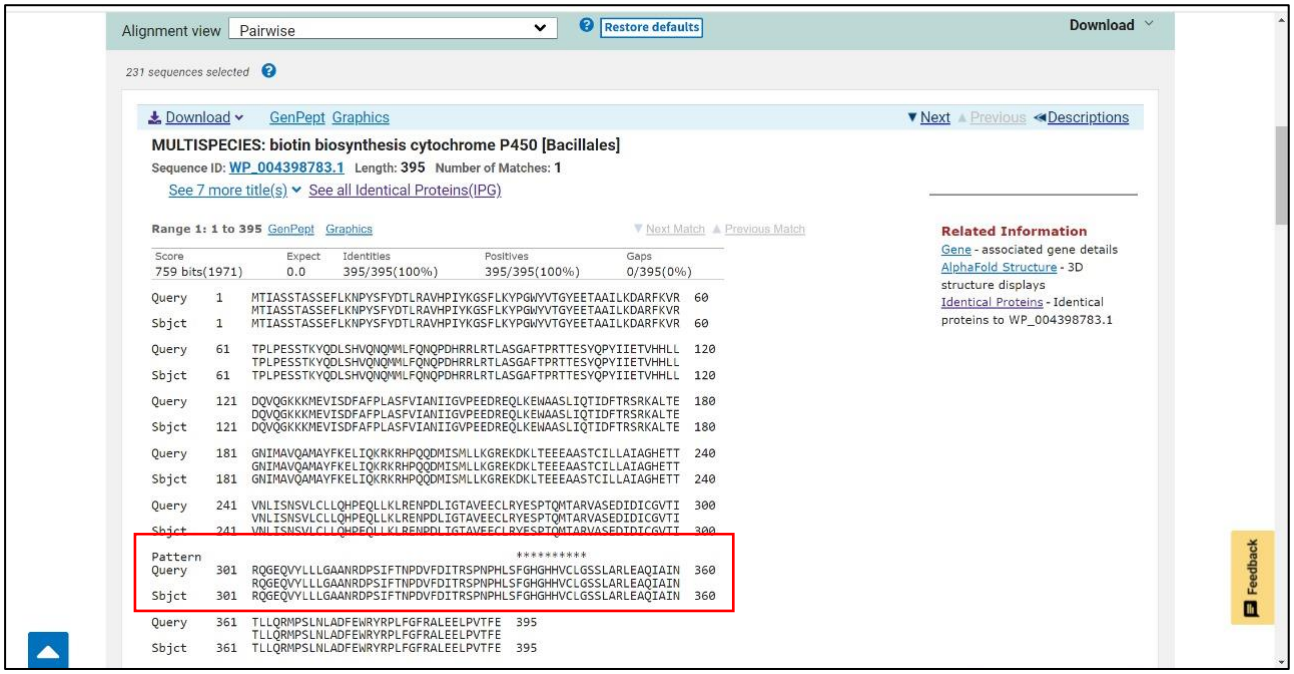


Figure 9: Result for Alignment Section

Descriptions | Graphic Summary | Alignments | Taxonomy

Reports: Lineage | Organism | Taxonomy

100 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
root			334	
synthetic construct	other sequences	1244	13	synthetic construct hits
Homo sapiens	primates	1239	236	Homo sapiens hits
Pongo abelii	primates	1239	5	Pongo abelii hits
Gorilla gorilla gorilla	primates	1229	1	Gorilla gorilla gorilla hits
Pan paniscus	primates	1228	1	Pan paniscus hits
Pan troglodytes	primates	1228	3	Pan troglodytes hits
Pongo pygmaeus	primates	1219	1	Pongo pygmaeus hits
Nomascus leucogenys	primates	1211	1	Nomascus leucogenys hits
Hylobates moloch	primates	1211	1	Hylobates moloch hits
Symphalangus syndactylus	primates	1206	1	Symphalangus syndactylus hits
unidentified	unclassified sequences	1188	2	unidentified hits
Macaca mulatta	primates	1175	4	Macaca mulatta hits
Macaca fascicularis	primates	1175	5	Macaca fascicularis hits
Macaca thibetana thibetana	primates	1174	1	Macaca thibetana thibetana hits
Theropithecus gelada	primates	1173	1	Theropithecus gelada hits
Macaca nemestrina	primates	1172	1	Macaca nemestrina hits

Feedback

Figure 10: Result for Taxonomy section based on 'Lineage'

Descriptions | Graphic Summary | Alignments | Taxonomy

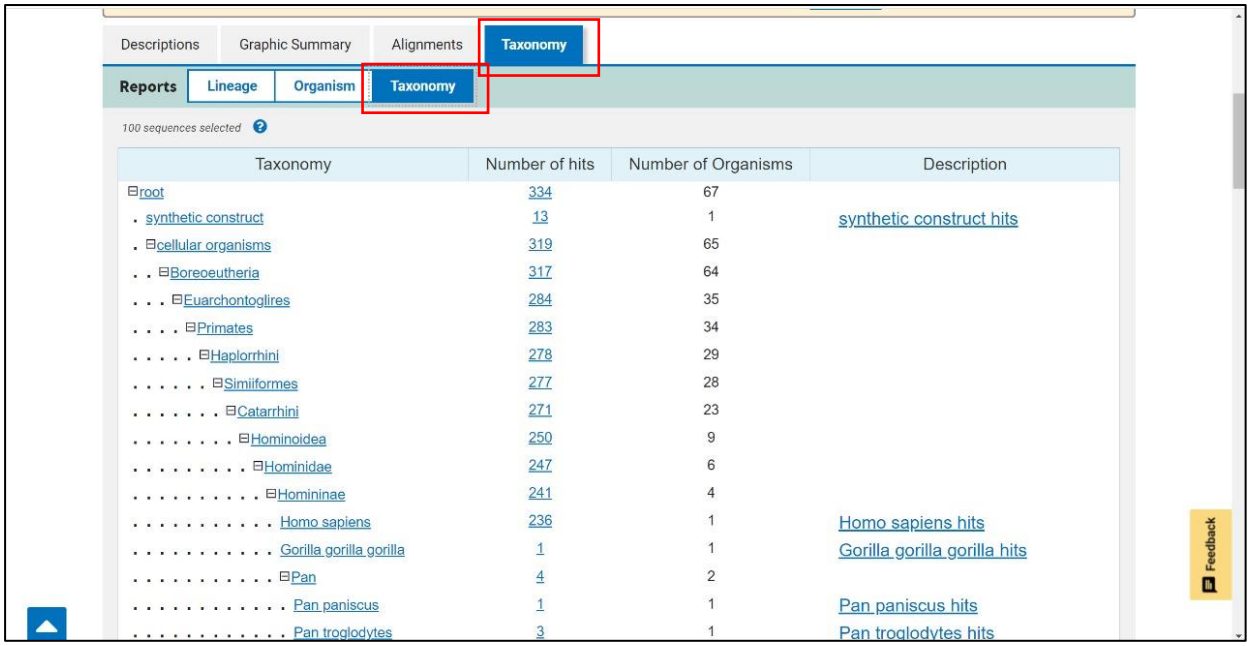
Reports: Lineage | Organism | Taxonomy

100 sequences selected

Description	Score	E value	Accession
synthetic construct [other sequences]	▼Next	▲Previous	◄◄First
serum albumin-interferon alpha 1 fusion protein, partial [synthetic construct]	1244	0.0	AGI02589
albumin, partial [synthetic construct]	1239	0.0	AAX36126
albumin [synthetic construct]	1239	0.0	ABM82340
serum albumin [synthetic construct]	1220	0.0	AIC32938
HSA-cIFN [synthetic construct]	1195	0.0	QCO95453
HSA-GGGGS-GH fusion protein, partial [synthetic construct]	1192	0.0	AFO84000
IL-1Ra-GGGGS-HSA fusion protein, partial [synthetic construct]	1191	0.0	AEL88488
HSA-GGGGS-IL-1Ra fusion protein, partial [synthetic construct]	1191	0.0	AEZ51871
human serum albumin and interferon-alpha2b fusion protein, partial [synthetic construct]	1190	0.0	QNI40628
HSA-GGGGS-PTH(1-34), partial [synthetic construct]	1189	0.0	AER13700
serum albumin, partial [synthetic construct]	1188	0.0	AIC32937
somatostatin (SST) doublet/albumin fusion protein [synthetic construct]	1186	0.0	UTT97830
human serum albumin mutein, partial [synthetic construct]	1185	0.0	QNI40627
Homo sapiens (human) [primates]	▼Next	▲Previous	◄◄First
albumin preproprotein [Homo sapiens]	1239	0.0	NP_000468
RecName: Full=Albumin; Flags: Precursor [Homo sapiens]	1239	0.0	P02768
Chain A, SERUM ALBUMIN [Homo sapiens]	1239	0.0	4BKE_A

Feedback

Figure 11: Result for Taxonomy section based on 'Organism'



Descriptions | Graphic Summary | Alignments | Taxonomy

Reports | Lineage | Organism | Taxonomy

100 sequences selected

Taxonomy	Number of hits	Number of Organisms	Description
⊟root	334	67	
. synthetic construct	13	1	synthetic construct hits
. ⊟cellular organisms	319	65	
. . ⊟Boreoeutheria	317	64	
. . . ⊟Euarchontoglires	284	35	
. . . . ⊟Primates	283	34	
. . . . . ⊟Haplorrhini	278	29	
. . . . . ⊟Simiiformes	277	28	
. . . . . ⊟Catarrhini	271	23	
. . . . . ⊟Hominoidea	250	9	
. . . . . ⊟Hominidae	247	6	
. . . . . ⊟Homininae	241	4	
. . . . . Homo sapiens	236	1	Homo sapiens hits
. . . . . Gorilla gorilla gorilla	1	1	Gorilla gorilla gorilla hits
. . . . . ⊟Pan	4	2	
. . . . . Pan paniscus	1	1	Pan paniscus hits
. . . . . Pan troglodytes	3	1	Pan troglodytes hits

Feedback

Figure 12: Result for Taxonomy section based on 'Taxonomy'

RESULTS:

Pattern-Hit Initiated BLAST (PHI-BLAST) tool is a variant of the Basic Local Alignment Search Tool (BLAST) algorithm, specifically designed for detecting distant relationships between protein sequences and identifying domains of potential functional significance within sequences. The tool was used to studied query where it is able to detect the pattern in the organisms which confirms the identification of remote homologs or conserved domains for the query protein sequences.

CONCLUSION:

PHI-BLAST is widely used in bioinformatics, particularly for analyzing protein sequences to identify conserved domains, motifs, or functional signatures. It aids in understanding evolutionary relationships between proteins and assists in annotating sequences with functional information based on conserved patterns. Its ability to focus the alignment and construction of the PSSM around a motif provides a valuable approach for researchers and bioinformaticians working in the field of protein analysis.

REFERENCES:

1. ResearchGate. (2023). BLAST Algorithm. https://www.reseatchgate.net/publication/230503487
2. Zheng Zhang, Webb Miller, Alejandro A. Schäffer, Thomas L. Madden, David J. Lipman, Eugene V. Koonin, Stephen F. Altschul, Protein sequence similarity searches using patterns as seeds, Nucleic Acids Research, Volume 26, Issue 17, 1 September 1998, Pages 3986–3990, https://doi.org/10.1093/nar/26.17.3986
3. Sancho J. Flavodoxins: sequence, folding, binding, function and beyond. Cell Mol Life Sci. 2006 Apr;63(7-8):855-64. doi: 10.1007/s00018-005-5514-4. PMID: 16465441. https://pubmed.ncbi.nlm.nih.gov/16465441

DATE: 01/11/23

WEBLEM 6(E)
EMBOSS NEEDLE – GLOBAL PAIRWISE SEQUENCE ALIGNMENT
(URL: https://www.ebi.ac.uk/Tools/psa/emboss_needle/)

AIM:

To explore and compare the protein sequences of 'Myosin' from two organisms *Gallus gallus* (UniProt ID: Q90623) and *Mus musculus* (UniProt ID: F8VQB6) by performing global pairwise sequence alignment using EMBOSS Needle Tool.

INTRODUCTION:

The European Molecular Biology Open Software Suite, or EMBOSS, is a part of the European Bioinformatics Institute (EBI). One of the prominent tools of EMBOSS is EMBOSS Needle, which is based on the Needleman-Wunsch algorithm. The Needleman-Wunsch algorithm was developed by Saul B. Needleman and Christian D. Wunsch in 1970 for global sequence alignment. It works on the principle of dividing the large problem into a series of smaller problems and uses the solutions to the smaller problems to find an optimal solution to the larger problem, assigning a score to every possible alignment and finding all possible alignments having the highest score.

The unique feature of the EMBOSS Needle tool is that it finds the alignment with the maximum possible score where the score of an alignment is equal to the sum of the matches taken from the scoring matrix, minus penalties arising from opening and extending gaps in the aligned sequences. The substitution matrix and gap opening and extension penalties are user-specified. A penalty is subtracted from the score for each gap opened (Gap insertion penalty) and a penalty is subtracted from the score for the extension of the inserted gaps (Gap extension penalty). Typically, the cost of extending a gap is set to be 5-10 times lower than the cost for opening a gap.

Penalty for a gap of n positions is calculated using the following formula:

Gap at n^{th} position = gap opening penalty + (n - 1) * gap extension penalty

Myosin:

Myosin is a motor protein with a primary role in muscle contraction, interacting with actin filaments to generate force and movement. Beyond muscles, myosin participates in cell motility, cell division, intracellular transport, and maintenance of cell shape, making it a crucial component in various cellular processes. The need to analyze myosin with the EMBOSS Needle tool arises from the diverse functions of myosin, which contribute to the dynamic behavior and structural integrity of cells. By analyzing the sequence and structure of myosin, researchers can gain insights into its mechanisms and interactions, which can help develop a deeper understanding of its role in various cellular processes and potentially lead to new therapeutic strategies for muscle and non-muscle related disorders.

METHODOLOGY:

1. Open the UniProt database and search for the query of 'Myosin'.
2. From the results page, open the proteins of interest. Here, *Gallus gallus* (UniProt ID: Q90623) and *Mus musculus* (UniProt ID: F8VQB6).
3. Download the myosin protein sequences of both the organisms in FASTA file format.
4. Open the homepage of EMBOSS Needle tool and paste the sequences in the query box and set the desired parameters. Select the 'SUBMIT' to submit the query.
5. The results page of EMBOSS Needle tool displays the Alignment, Submission Details and View Alignment File. Interpret the results.

OBSERVATIONS:

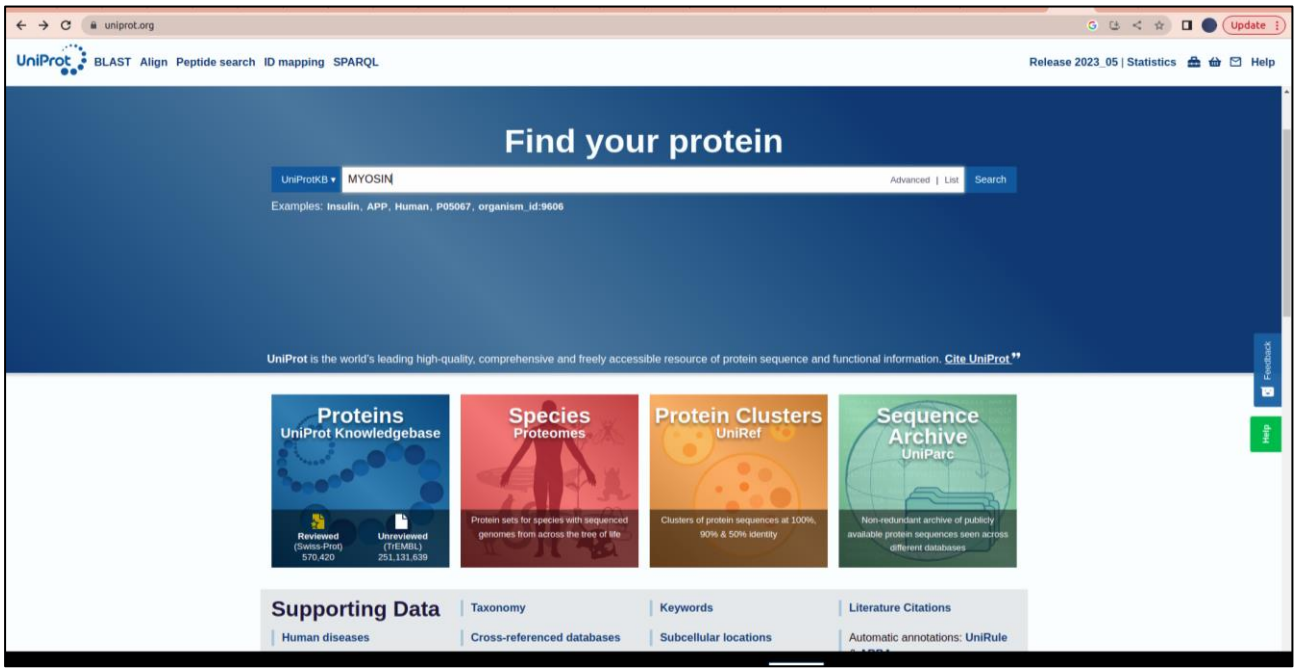


Figure 1: Homepage of the UniProt Database

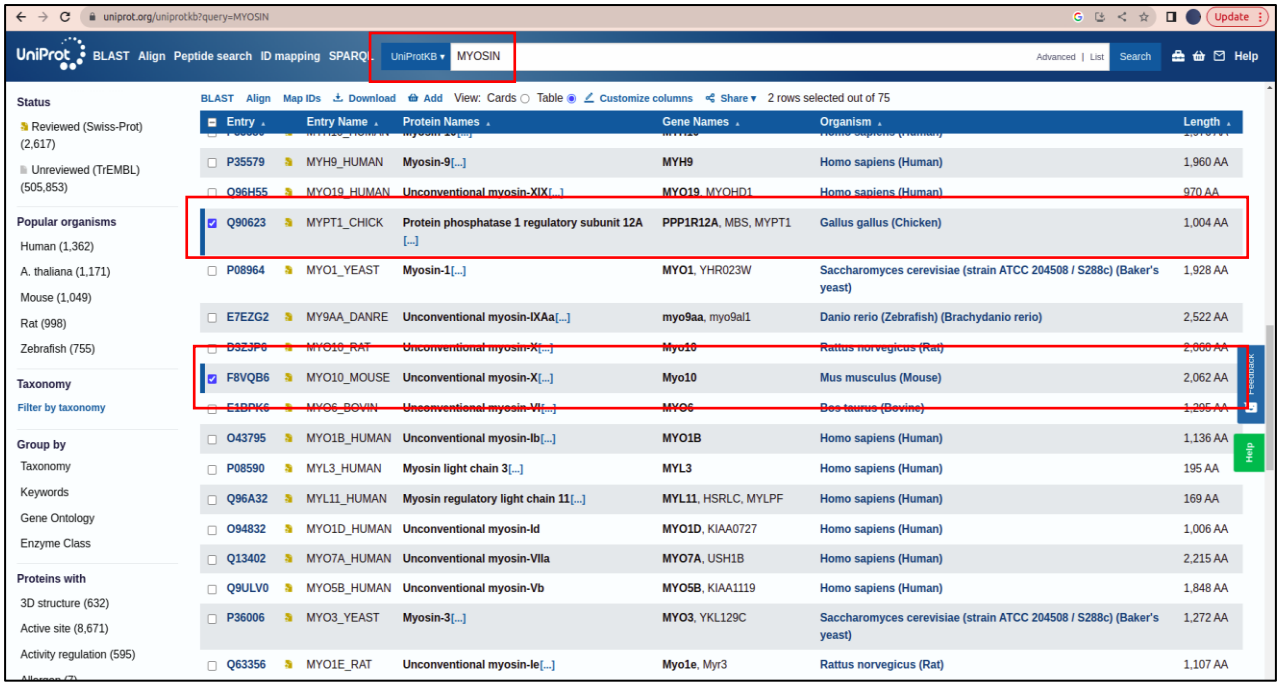


Figure 2: Results page of the UniProt Database for the query of Myosin with selected entries

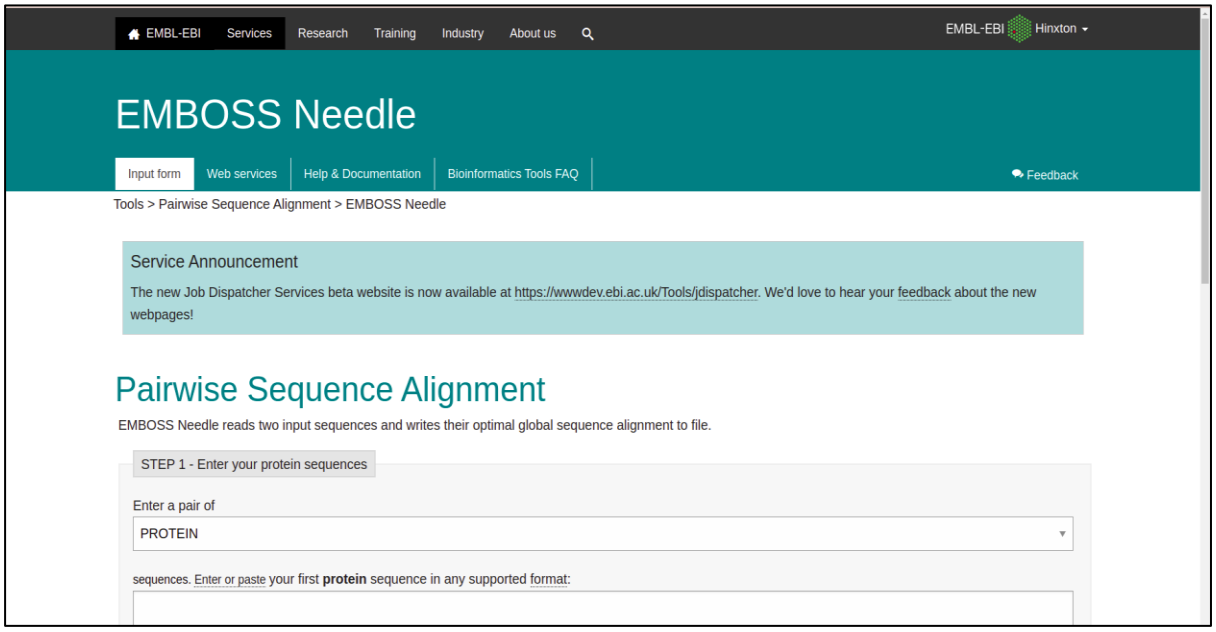


Figure 3: Homepage of EMBOSS Needle Tool

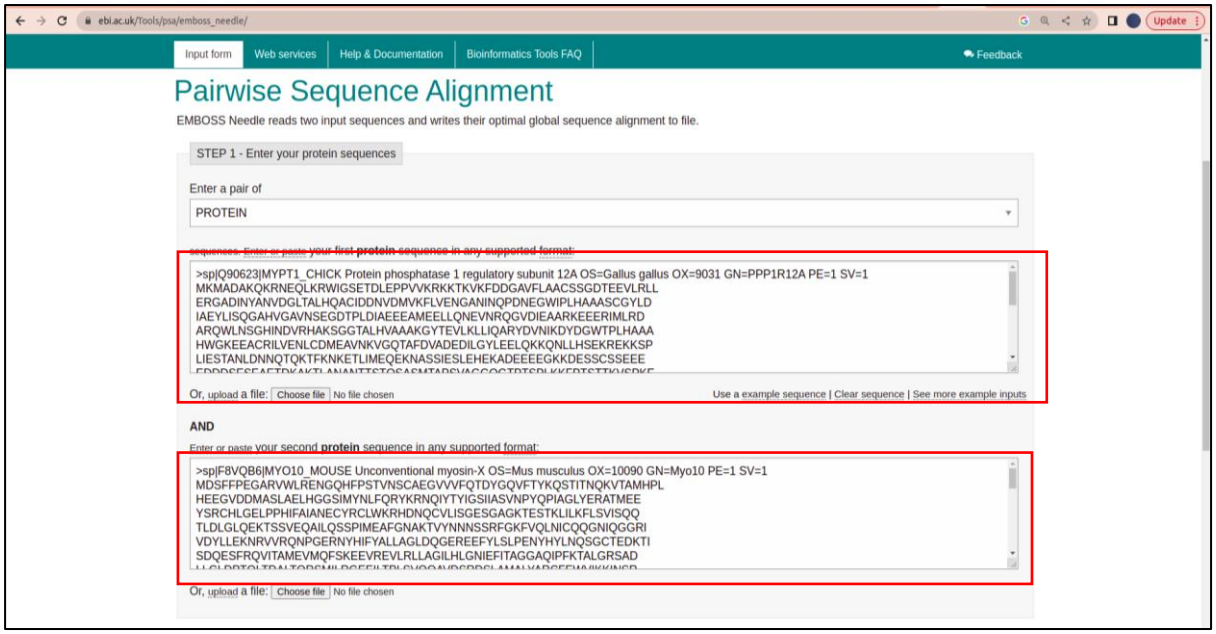


Figure 4: Submission of the protein sequences retrieved from the UniProt Database in the EMBOSS Needle Tool

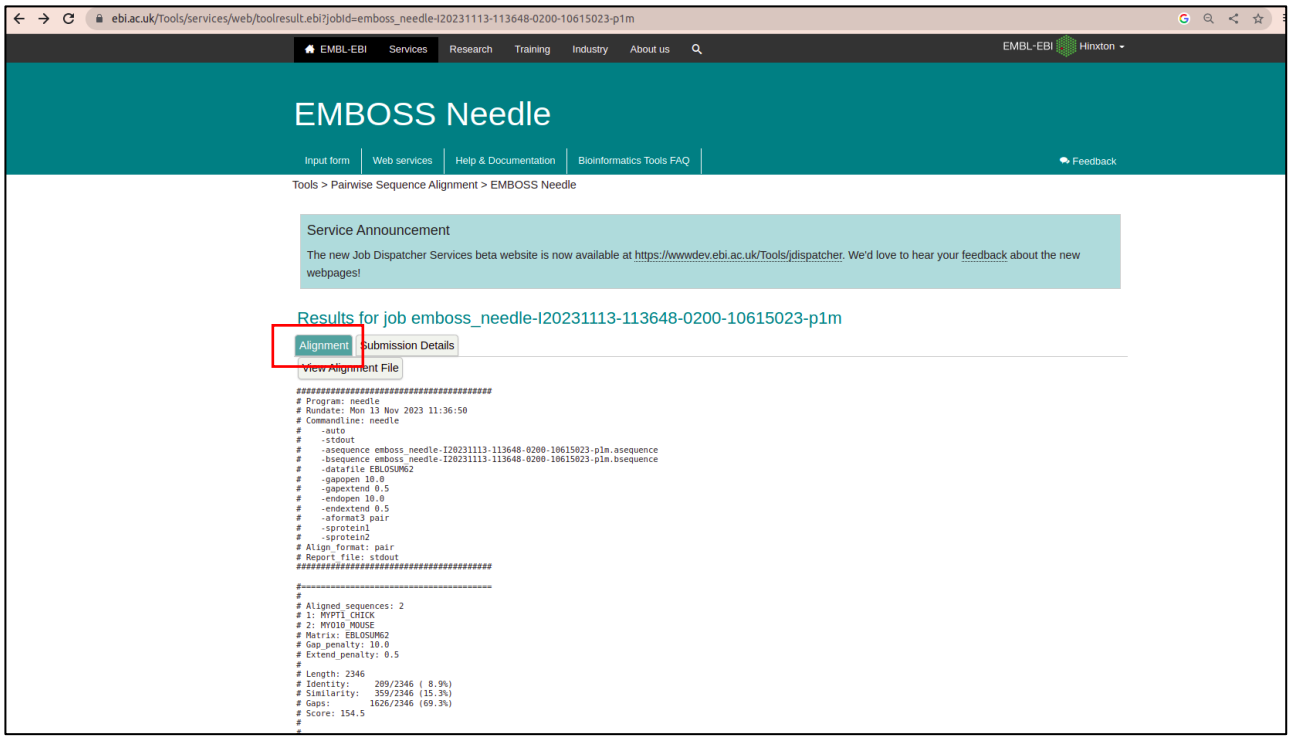


Figure 5: Results page of the submitted query with Alignment option

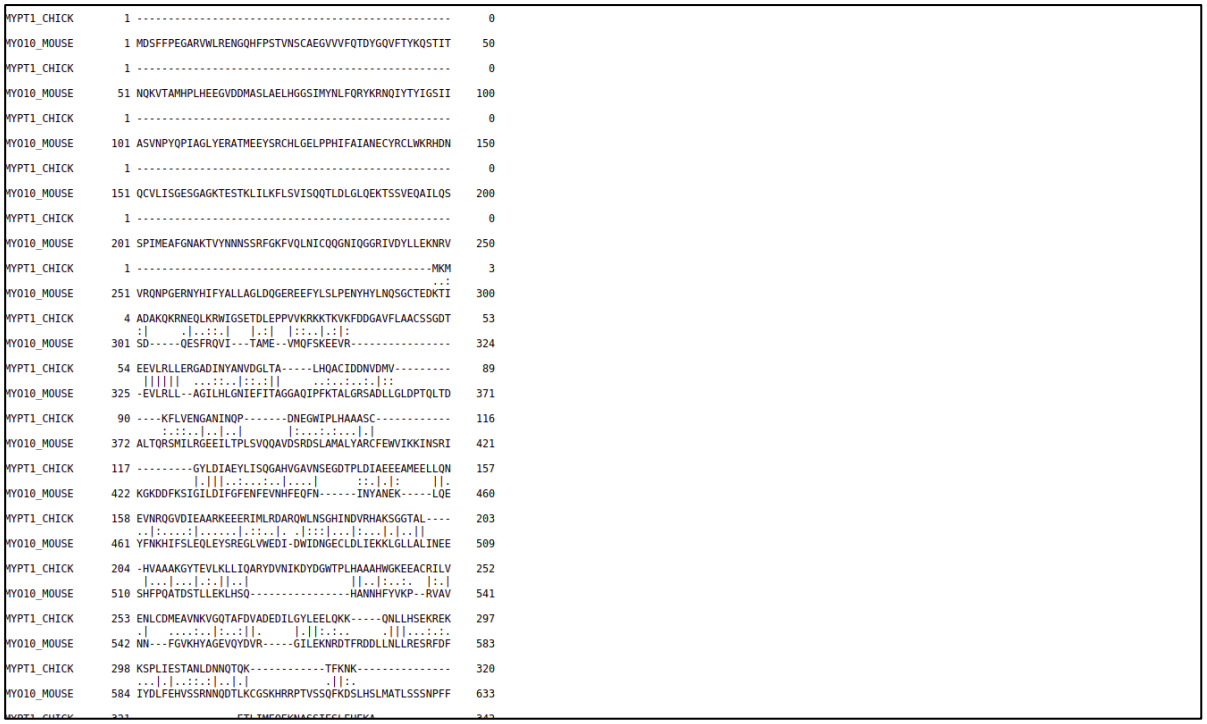


Figure 5a: Results page of the submitted query with Alignment option

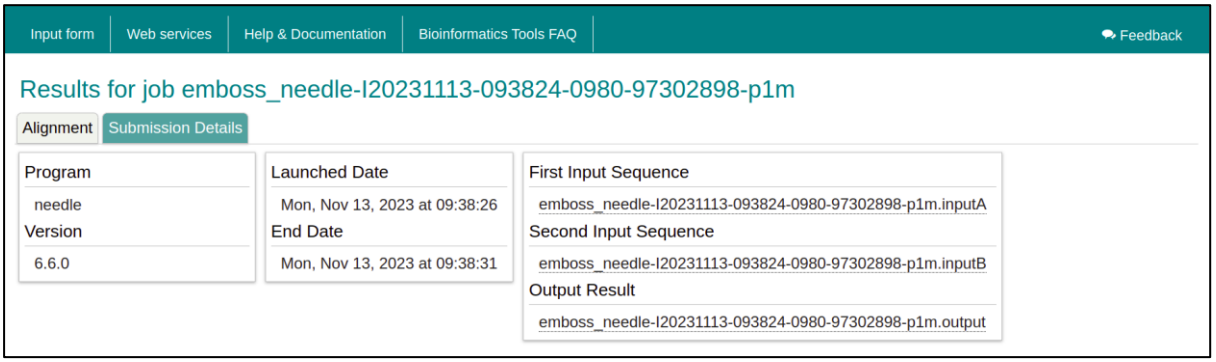


Figure 6: View of submission details

RESULTS:

By exploring global pairwise sequence alignment using the EMBOSS Needle tool, the results were observed and studied for the protein query 'Myosin' in organisms *Gallus gallus* (UniProt ID: Q90623) and *Mus musculus* (UniProt ID: F8VQB6). It was found that in the pairwise alignment of the two organisms, they were not identical upon comparison, as the sequence identity is only 8.9%.

Length	2346
Identity	209/2346 (8.9%)
Similarity	359/2346 (15.3%)
Gaps	1626/2346 (69.3%)
Score	154.5

CONCLUSION:

EMBOSS Needle tool, for Global Pairwise Sequence Alignment, was explored by comparative study of protein 'Myosin' of two different organisms, namely, *Gallus gallus* (UniProt ID: Q90623) and *Mus musculus* (UniProt ID: F8VQB6).

REFERENCES:

1. Needleman, S. B. and Wunsch, C. D. (1970) *J. Mol. Biol*. 48, 443-453. https://www.bioinformatics.nl/cgi-bin/emboss/help/needle
2. Robert S. Adelstein, James R. Sellers, in *Biochemistry of Smooth Muscle Contraction*, 1996. https://doi.org/10.1016/B978-0-12-801387-8.00003-X

DATE: 01/11/23

WEBLEM 6(F)
EMBOSS WATER – LOCAL PAIRWISE SEQUENCE ALIGNMENT
(URL: https://www.ebi.ac.uk/Tools/psa/emboss_water/)

AIM:

To explore and compare the protein sequences of 'collagen' in two organisms, *Rattus norvegicus* (UniProt ID: P05539) and *Homo sapiens* (UniProt ID: P08572), by performing local pairwise sequence alignment using the EMBOSS Water tool.

INTRODUCTION:

The European Molecular Biology Open Software Suite, or EMBOSS, is a part of the European Bioinformatics Institute (EBI). One of the prominent tools of EMBOSS is EMBOSS Water, which is based on the Smith-Waterman algorithm. Smith-Waterman algorithm was developed by Temple F. Smith and Michael S. Waterman in 1981 and is used for local sequence alignment, which finds the best subsequence match between two sequences by comparing all possible pairs of subsequences. The unique aspect of the EMBOSS Water tool is that it uses a speed-accelerated version of the Smith-Waterman method to determine the local alignment of a sequence with one or more other sequences. By examining every potential alignment and choosing the best one, dynamic programming techniques guarantee the best possible local alignment. To do this, a scoring matrix with values for each potential residue or nucleotide match is incorporated.

The EMBOSS Water tool employs a modified Smith-Waterman algorithm with speed enhancements to compute the local alignment of one or more sequences. Users have the flexibility to specify the gap insertion penalty, gap extension penalty, and substitution matrix for calculating alignments. The output is a standard EMBOSS alignment file. Identity refers to the percentage of identical matches between two sequences over the entire reported aligned region, inclusive of any length gaps. Similarly, similarity represents the percentage of matches between the two sequences over the length of the reported aligned region, considering any gaps.

Collagen:

The most prevalent protein in the body, collagen, is found in various connective tissues such as the skin, tendons, bones, and ligaments. Its inherent stiffness and resistance to stretching contribute significantly to providing structural support within the extracellular space of connective tissues. Understanding collagen's structure, function, and its implications in various diseases and conditions, including autoimmune disorders like rheumatoid arthritis, lupus, dermatomyositis, and scleroderma, is crucial. These conditions can adversely affect collagen, highlighting the importance of in-depth research.

The EMBOSS Water tool serves as a valuable resource in this pursuit. It is a pairwise sequence alignment program designed to determine the local alignment of one or more sequences. The tool utilizes a modified version of the Smith-Waterman technique, offering faster results for

researchers. By employing the EMBOSS Water tool to analyze collagen, researchers can gain deeper insights into its molecular makeup and its role in health and disease.

METHODOLOGY:

1. Open the UniProt database and search for the query of 'Collagen'.
2. From the results page, open the proteins of interest. Here, *Rattus norvegicus* (UniProt ID: P05539) and *Homo sapiens* (UniProt ID: P08572).
3. Download the collagen protein sequences of both the organisms in FASTA canonical file format.
4. Open the homepage of EMBOSS Water tool and paste the sequences in the query box and set the desired parameters. Select the 'SUBMIT' to submit the query.
5. The results page of EMBOSS Water tool displays the Alignment, Submission Details and View Alignment File. Interpret the results.

OBSERVATIONS:

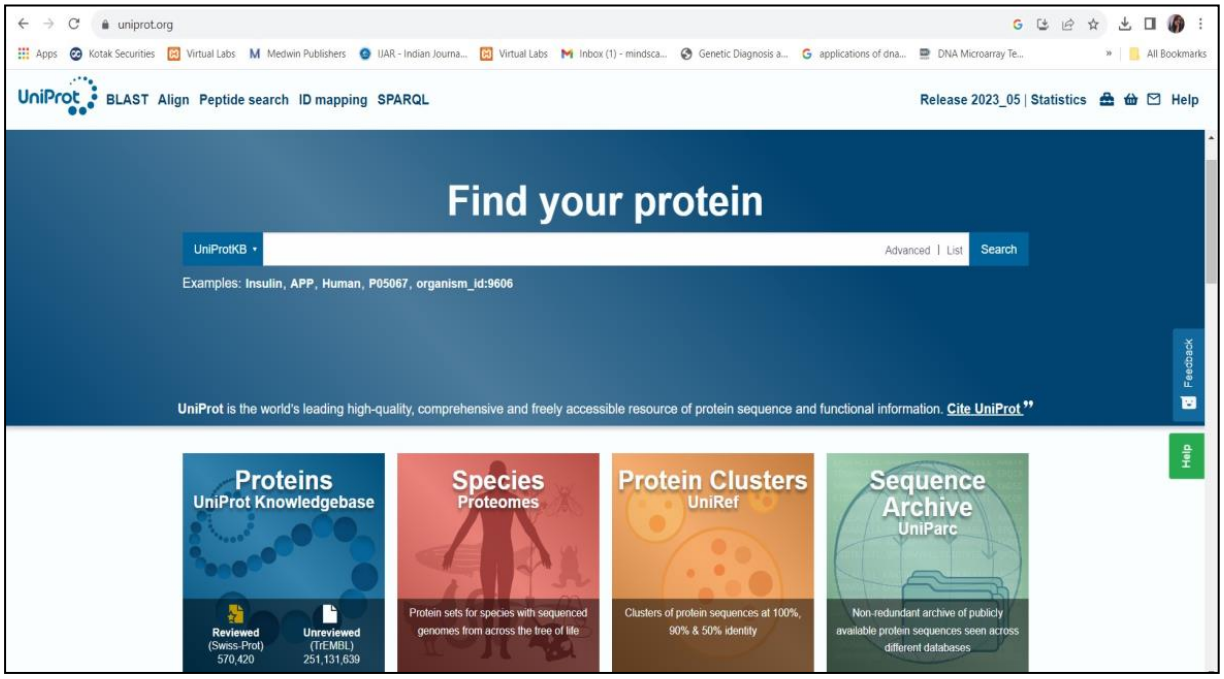


Figure 1: Homepage of the UniProt database

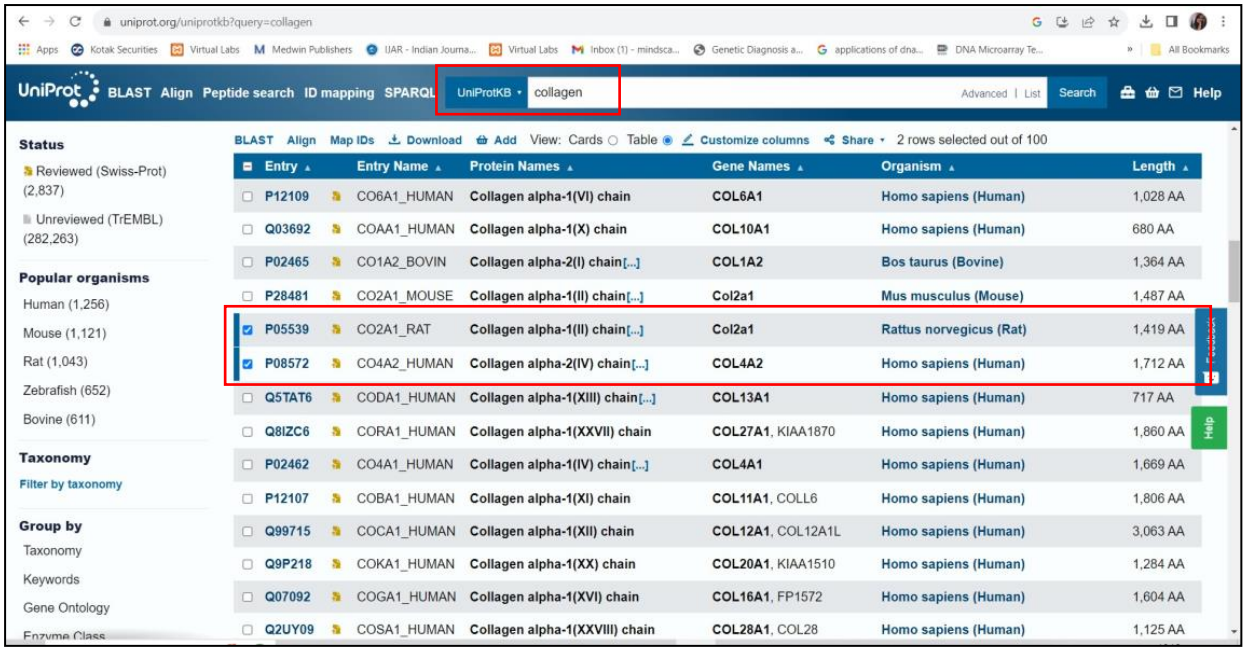


Figure 2: Results page of the UniProt Database for the query of collagen with selected entries

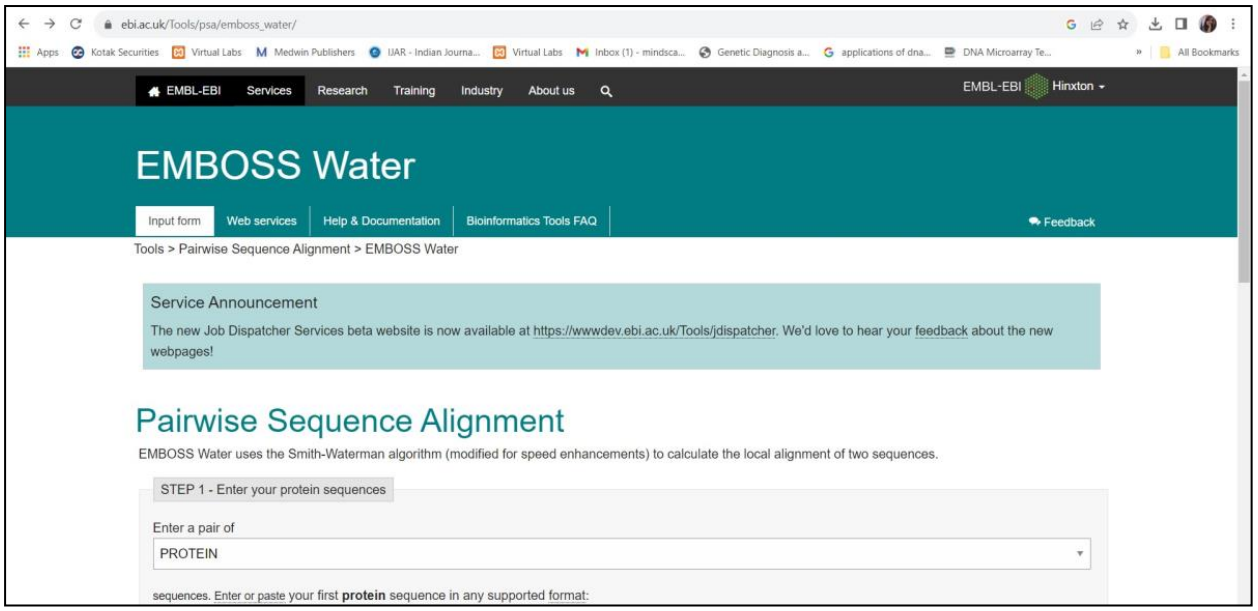


Figure 3: Homepage of EMBOSS Water Tool

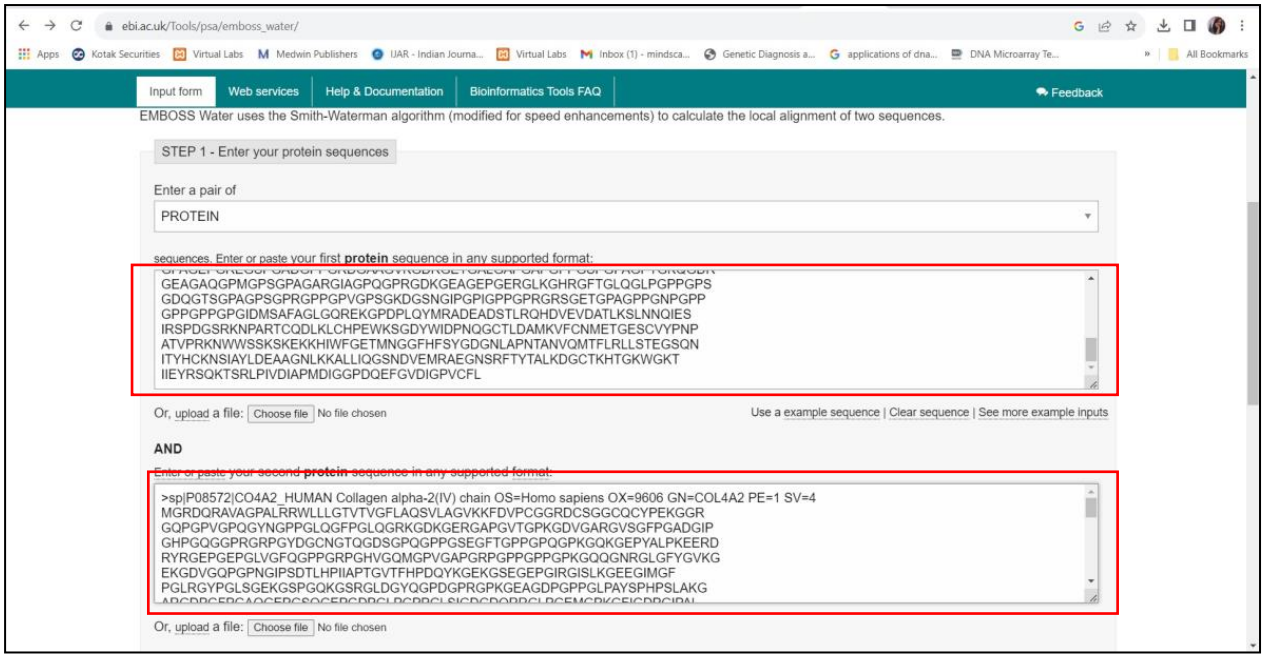


Figure 4: Submission of the protein sequences retrieved from the UniProt Database in the EMBOSS Water Tool

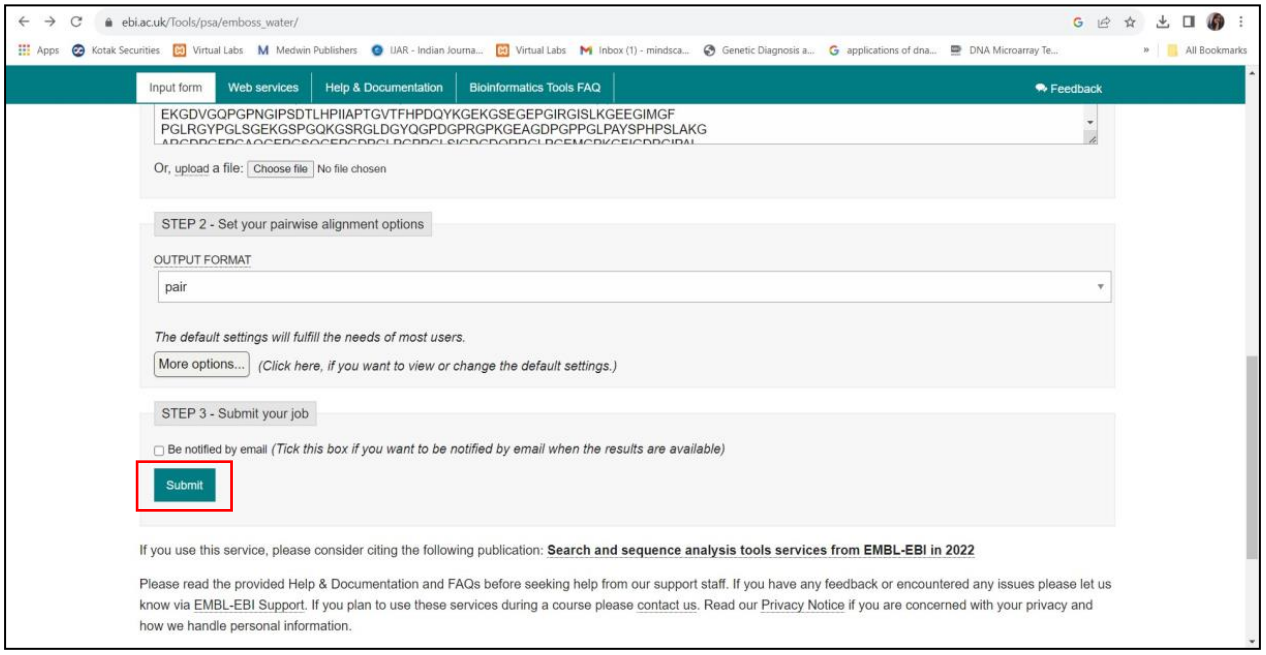


Figure 5: Submission of the query to the EMBOSS Water Tool

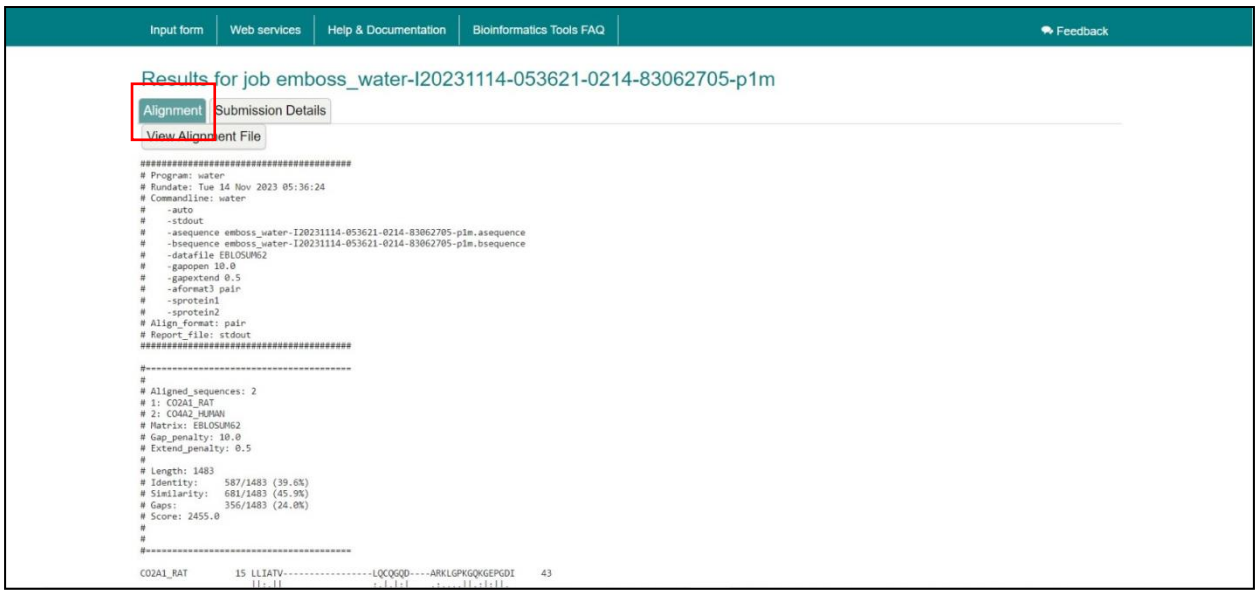


Figure 6: Results page of the submitted query with Alignment option

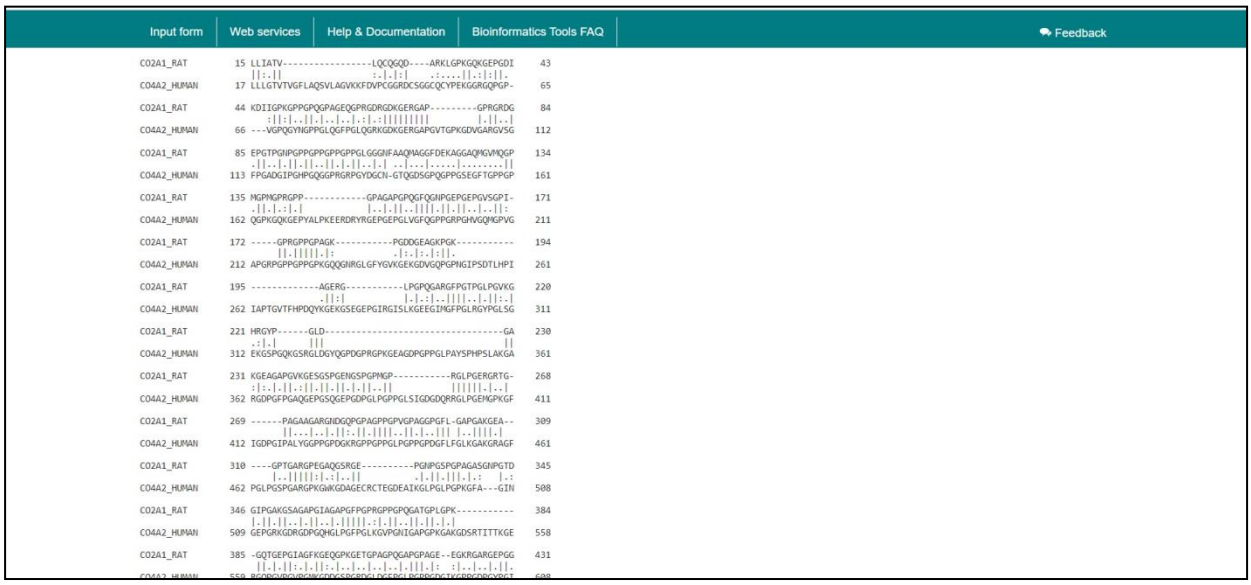


Figure 6a: Results page of the submitted query with Alignment option

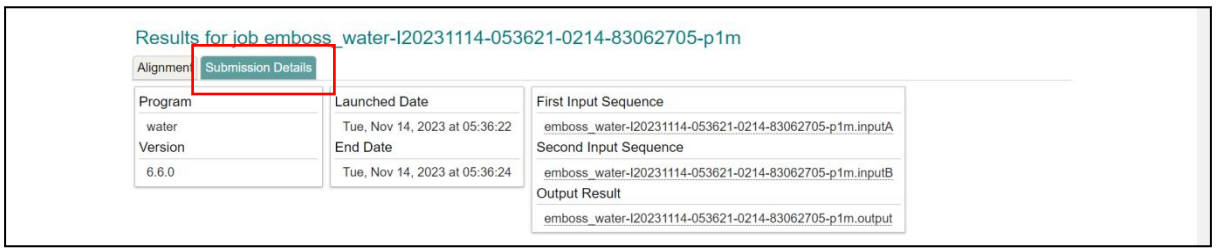


Figure 7: View of Submission details

RESULTS:

By exploring local pairwise sequence alignment using EMBOSS Water Tool, the results were observed and studied for query for protein query 'collagen' for organism *Rattus norvegicus* (UniProt ID: P05539) and *Homo sapiens* (UniProt ID: P08572) and it was observed that the local pairwise sequence alignments of the two organisms were found to be identical upon comparison, as the sequence identity is 39.6%.

Length	1483
Identity	587/14683 (39.6%)
Similarity	681/1483 (45.9%)
Gaps	356/1483 (69.3%)
Score	2455.0

CONCLUSION:

EMBOSS Water tool, for Local Pairwise Sequence Alignment, was explored by comparative study of protein collagen of two different organisms, namely, *Rattus norvegicus* (UniProt ID: P05539) and *Homo sapiens* (UniProt ID: P08572).

REFERENCES:

1. Smith TF, Waterman MS (1981) *J. Mol. Biol* 147(1). https://emboss.sourceforge.net/apps/release/6.6/emboss/apps/water.html
2. H. Jawad, R.A. Brown, in *Comprehensive Biotechnology*, 2011. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/collagen