PROJECT REPORT

**To find the protein/proteins present in the least number of organisms**

****

**Group Members**

Shaik Rumaan - BIM2016004

Utkarsh Singh - MBI2019011

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

**UNDER THE SUPERVISION OF**

DR. BS SANJEEV

# 

# **TABLE OF CONTENTS**

1. Introduction 04

2. Problem statement 05

3. Dataset 05

4. Approach 08

5. Result 11

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# **Abstraction**

The comparison of nucleotide or protein sequences from the same or different organisms is a very powerful tool in molecular biology. By finding similarities between sequences, scientists can infer the function of newly sequenced genes, predict new members of gene families, and explore evolutionary relationships. Now that whole genomes are being sequenced, sequence similarity searching can be used to predict the location and function of protein-coding and transcription regulation regions in genomic DNA Sequence alignment becomes challenging with an increase in size and number of sequences.Finding optimal or near optimal solutions for sequence alignment is one of the most important operations in bioinformatics. To handle this situation an algorithm developed in 1990 called BLAST (Basic Local Alignment Search Tool) is a sequence similarity search tool. BLAST compare all combinations of nucleotide or protein queries with nucleotide or protein databases.

# 

# **INTRODUCTION**

Sequence similarity searching to identify homologous sequences is one of the first, and most informative, steps in any analysis of newly determined sequences. Modern protein sequence databases are very comprehensive, so that more than most of metagenomic sequence samples typically share significant similarity with proteins in sequence databases as like UniProt

UniProt contains high-quality manually annotated and non-redundant protein sequence records. Manual annotation consists of analysis, comparison and merging of all available sequences for a given protein, as well as a critical review of associated experimental and predicted data. UniProt curators extract biological information from the literature and perform numerous computational analyses.

An approach to rapid sequence comparison, basic local alignment search tool (BLAST),directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates .Blast is the tool most frequently used for calculating sequence similarity. BLAST comes in variations for use with different query sequences against different databases.as output bast introduces some parameters max target sequences,Word size,Max matches in a query range,Match/mismatch scores,Gap Costs ,E-value to calculate sequence similarity.

# 

# 

# 

# **PROBLEM DEFINITION**

Fetch out the proteomes of 10 organisms and find out the Protein /Proteins that is present in least number of organism.

**Given** : 10 organisms with there Proteins name and Protein sequences

**Output required**: Find out the Protein names that is present in least number of organism

# **DATABASE**

The dataset containing proteome of 10 different organisms was downloaded from UniProt database. They are in FASTA file format. The organisms whose proteome has been downloaded are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **SR.**  **NO.** | **ORGANISM NAME** | **KINGDOM** | **TOTAL PROTEINS** |
| 1. | Clostridium sp. CAG:138 | Bacteria | 1746 |
| 2. | Bovine papillomavirus type 5 | Virus | 7 |
| 3. | Mycoplasma caviae | Bacteria | 1008 |
| 4. | Escherichia phage EcWhh-1 | Virus | 279 |
| 5. | Sulfolobus acidocaldarius (strain ATCC 33909) | Archae | 2221 |
| 6. | Salmonella virus PsP3 | Virus | 42 |
| 7. | Thermococcus onnurineus  (strain NA1) | Archae | 1976 |
| 8. | Nematocida parisii (strain ERTm3) | Eukaryote | 2724 |
| 9. | Methanofollis liminatans DSM 4140 | Archae | 2423 |
| 10. | Penicillium digitatum (strain PHI26 / CECT 20796) | Eukaryote | 9101 |

## **Universal Protein Resourc**e **(UniProt):**

UniProt is comprehensive resource for protein sequence and annotation data. It is freely accessible database of protein sequence and functional information, entries being derived from genome sequencing projects. It contains a large amount of information about the biological function of proteins derived from research literature. It is a central repository of protein data created by combining the Swiss-Prot TrEMBL and PIR-PSD databases.

The UniProt consortium comprises the European Bioinformatics institute (EBI), the Swiss institute of Bioinformatics(SIB), and the Protein information Resource(PIR). EBI, located at the wellcome Trust Genome Campus in Hinxton, UK,hosts a large resource of bioinformatics databases and services.

## **UniProtKB/Swiss-Prot:**

UniProtKB/Swiss-Prot is a manually annotated non-redundant protein sequence database. It combines information extracted from scientific literature and biocurator-evaluated computational analysis. The aim of UniProtKB/Swiss-Prot is to Provide all known relevant information about a particular protein.

# **APPROACH**

* Fetch out the Proteomes of organisms from Uniprot database.
* we are using blast to find the sequence similarity
* when we run blast we get an output file which has ids of protein and corresponding protein id that is similar to given query sequence with e value of 10.
* we will maintain a corresponding protein id and count of protein id having similar sequences by traversing the files generated from blast.
* we will maintain a corresponding protein id and protein name for all the organism.
* if we found any such protein ids from the organisms which are not having similarity with any sequence so these are the proteins that are unique so these are the proteins that are present in least number of organisms
* else we compare the count of protein ids and take the count of ids which are minimum.Then we display the protein names which are present in least number of organisms

# 

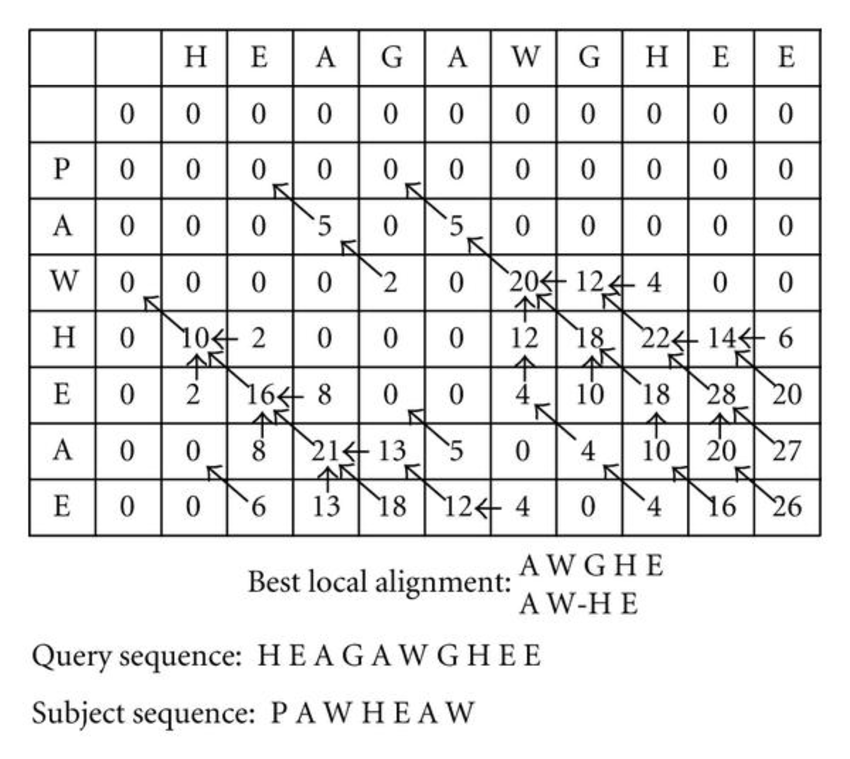
# 

# 

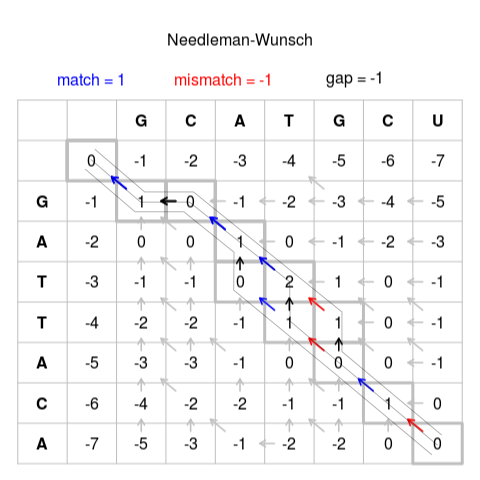
# 

**ALIGNMENT**

Alignment is Arrangement in correct relative positions in bioinformatics we use sequence alignment Sequence Alignment is a method of arranging the sequence of DNA RNA and Protein to identify regions of similarity that may be a consequences Alignment sequences of nucleotide or amino acid residues are typically represented as rows within a matrix.

**Local Alignment:** Local Alignment technique is the Smith-Waterman algorithm, which is based on dynamic programming with additional choices to start and end at any place. In local alignment as result we get multiply alignments. ****

**Global Alignment:** Global Alignment technique is the Needleman-Wunsch algorithm, which is based on dynamic programming. In Global Alignment as result we get single alignment



**Blast:**

Blast is Basic Common Local Alignment Search Tool which introduced in in 1990 by

Stiphen Altschul.Blast is use to compare gene and protein sequences.It is set of sequence comparison algorithm use to search databases for optimal local alignment to a query. It breaks query and data bases sequences into fragments And search matches in between them.protein alignment is time consuming by Dynamic programming. Blast is 50 times faster than dynamic programming.

How Blast Works:

* Blast Searches for short regions of a given length(w) called words That score at least (T) when compared to the query sequence that align with sequences in the database **target sequences** using a substitution matrix.
* For every pair of sequences that have a words in common, blast extends the alignment

In both direction to find alignments that score greater than a certain score threshold

(S) these alignments are called high scoring pair or HSPs,the maximum scoring pair HSPs are called MSPs.

E- value(expectation value):

* Expectation value is a parameter which describes the numbers of hits which we can expect.
* It decreases exponentially when score (S) of match increases which means smaller

E-value more likely the match in significant.

**Results**

Proteins that are present in least number organism are two they are:

ID: tr|K9FLM7|K9FLM7\_PEND2 Protein: Uncharacterized protein

ID: tr|K9G6V6|K9G6V6\_PEND2 Protein: Uncharacterized protein

Number organism it is present is :1

ID: tr|K9FLM7|K9FLM7\_PEND2

Organism Name:Penicillium digitatum (strain PHI26 / CECT 20796)

Protein: Uncharacterized protein

Sequence:

MSFLRHAHPSNGLPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPXPPPPT

Sequence length:161

Function:protein with accession id :K9FLM7, the protein with which it showed the highest similarity was ATP-dependent DNA helicase RecQ.The function of this protein is that it initiates and disrupts DNA recombination at the time of DNA replication.

ID: tr|K9G6V6|K9G6V6\_PEND2

Organism Name:Penicillium digitatum (strain PHI26 / CECT 20796)

Protein: Uncharacterized protein

Sequence:

MTQCSAWRIEKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKQGKKKKIDKPASSTDVAFC

Sequence length:216

Function:protein with accession id :K9G6V6 , the protein with which it showed the highest similarity was Retrovirus-related Pol polyprotein from type-1 retrotransposable element R2 [Symbiodinium microadriaticum] with the lowest E-value (0.056).

The function of this protein is that it shows catalytic activity in the DNA conversion.

**Interpretation of results:**

We performed individual BLAST of each of the 2 sequences that we obtained as unique proteins.The BLAST was performed on the online protein BLAST tool.

For the protein with accession id :K9G6V6 , the protein with which it showed the highest similarity was Retrovirus-related Pol polyprotein from type-1 retrotransposable element R2 [Symbiodinium microadriaticum] with the lowest E-value (0.056).The function of this protein is that it shows catalytic activity in the DNA conversion.

protein with accession id :K9FLM7, the protein with which it showed the highest similarity was ATP-dependent DNA helicase RecQ.The function of this protein is that it initiates and disrupts DNA recombination at the time of DNA replication.

**References**

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment SearchTool.J Mol Biol 1990;215:403–410. [PubMed: 2231712]

2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389– 3402. [PubMed: 9254694]146917

3.The BLAST Sequence Analysis Tool[Chapter 16]Tom Maidden

4.Alignment free similarity analysis for protein sequences based on fuzzy integral(Ajay Kumar Shaw,Binod Chandra Tripathi and Soumyadeep Nand