**CLASSIFICATION OF BACTERIA STRAINS THROUGH THE ANALYSIS OF PROTEIN SEQUENCES**

SUBMITTED BY

MUHAMMAD ZOHAIB KHAN

SP21-MSCS-0028

SUPERVISED BY

Dr. SHAUKAT WASI

CO-SUPERVISOR

Ms. SUMRA KHAN

Logo, company name

Description automatically generated

*THESIS SUBMITTED TO THE FACULTY OF COMPUTING)], MOHAMMAD ALI JINNAH UNIVERSITY, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE*

SPRING 2022

**Certificate of Approval**

****

*It is certified that the research work presented in this thesis, titled Classification of Bacteria Strains Through the Analysis of Protein Sequences was conducted by [Muhammad Zohaib Khan under the supervision of Dr. Shaukat Wasi.*

*No part of this thesis has been submitted anywhere else for any other degree.*

*This thesis is submitted to the Department of Computer Science in partial fulfilment of the requirements for the degree of*

*Master of Science in Computer Science in Spring 2022*

*at the*

*Mohammad Ali Jinnah University*

*Karachi, Pakistan*

*[Place Date of Thesis Defense here (for example Month DD, YYYY)]*

Candidate Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Signature: **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Examination Committee:**

1. Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Coordinator, HoD, Dean,

**Certificate of SUPERVISION**

****

*This is to certify that the thesis titled, “Classification of Bacteria Strains Through the Analysis Of Protein Sequences”, is submitted to the Department of Computer Science, Spring 2021, by Muhammad Zohaib Khan for the award of the degree of Master of Science in the discipline of Computer Science. The thesis has been carried out under my supervision. I certify that the work submitted is original and not plagiarized from any other source, except as specified in the references. Neither the thesis nor the work contained therein has been previously submitted to any other institution for a degree.*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Co-Supervisor’s Name and Signature Supervisor’s Name and Signature

**MOHAMMAD ALI JINNAH UNIVERSITY**

# ORIGINAL LITERARY WORK DECLARATION

|  |  |  |
| --- | --- | --- |
| **Faculty** | *Computing* | |
| **Program** | *MS Computer Science* | |
| Student Name: Muhammad Zohaib Khan | | Reg. No: sp21-mscs-0028 |
| Email: [sp21mscs0028@maju.edu.pk](mailto:sp21mscs0028@maju.edu.pk) | | Mobile No: 03312554474 |

Research Title:

|  |
| --- |
| CLASSIFICATION OF BACTERIA STRAINS THROUGH THE ANALYSIS OF PROTEIN SEQUENCES  I do solemnly and sincerely declare that:  1. I am the author of this work.  2. This work is original.  3. Use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work; I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work.  4. I hereby assign all and every rights in the copyright to this work to Mohammad Ali Jinnah University (MAJU), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of MAJU having been first had and obtained.  5. I am fully aware that if in the course of making this work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by MAJU. |

Student’s Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Place: Karachi-Pakistan.

**Acknowledgements**

I am thankful to Almighty for giving me the strength to work on a topic that I have never been through. I am deeply thankful to my supervisor Dr. Shaukat Wasi for his guidance, support and knowledge sharing. It was hardly possible to proceed. I am thankful to my family for understanding my tough timelines and supporting me

**Abstract**

Human efforts alone are insufficient to decipher the significance of protein sequences. The advancement of artificial intelligence has given humans the ability to design algorithms that can better interpret natural language. Due to the huge amount of data contained in biological patterns, determining their meaning is usually difficult. Nature's language is tough to interpret for the common person. Natural language and protein sequence are highly similar because both are character-based languages. This research work focuses on the classification of bacterial strains through the analysis of Protein Sequence. The name of the bacteria is stentrophomonas\_maltophilia. In this study we use different NLP techniques to extract characteristics and give appropriate representations for protein sequences. This research work is divided into 3 phases. Strain sequence collection, Signature identification of each strain and Classification using local alignment

Different NLP techniques can be applied on the given data set.  One of the techniques used for testing purpose are k-mer analysis using Jaccard Similarity and Bloom Filter.

*Keywords*: [Natural Language Processing, Bio Informatics, Classification]

Table of Contents

[ORIGINAL LITERARY WORK DECLARATION iv](#_Toc104876395)

[CHAPTER 1 4](#_Toc104876396)

[INTRODUCTION 4](#_Toc104876397)

[MOTIVATION 5](#_Toc104876398)

[SIGNIFICANCE 5](#_Toc104876399)

[CHAPTER 2 6](#_Toc104876400)

[LITERATURE REVIEW 6](#_Toc104876401)

[2.1 MATHFEATURE: FEATURE EXTRACTION PACKAGE FOR BIOLOGICAL SEQUENCES BASED ON MATHEMATICAL DESCRIPTORS 6](#_Toc104876402)

[2.2 BASIC LOCAL ALIGNMENT SEARCH TOOL 7](#_Toc104876403)

[2.3 EFFICIENT COUNTING OF K-MERS IN DNA SEQUENCES USING A BLOOM FILTER 8](#_Toc104876404)

[2.4 SPECTRAL JACCARD SIMILARITY: A NEW APPROACH TO ESTIMATING PAIRWISE SEQUENCE ALIGNMENTS 9](#_Toc104876405)

[EXPERIMENTAL SECTION 10](#_Toc104876406)

[MATH FEATURE 10](#_Toc104876407)

[PLAGIARISM VERIFICATION (for Final Submission) 14](#_Toc104876408)

[PLAGIARISM REPORT 0](#_Toc104876409)

[**CHECKLIST FOR FORMAT OF THESIS REPORTS** 3](#_Toc104876410)

CLASSIFICATION OF BACTERIA STRAINS THROUGH THE ANALYSIS OF PROTEIN SEQUENCE

MUHAMMAD ZOHAIB KHAN

Mohammad Ali Jinnah University

Author Note

Department of Computer Science, 22-E, Block-6, P.E.C.H.S., Lal Kothi Stop, Main Shahrah-e-Faisal, Karachi, 75400, Sindh, Pakistan. Email ID: [sp21mscs0028@maju.edu.pk](mailto:sp21mscs0028@maju.edu.pk)

# CHAPTER 1

## INTRODUCTION

Protein sequence is a combination of 20 different amino acids. Each amino acid has its own short code letter (Bonidia et al., 2020). Because of the massive amount of data available, labelling the data is difficult which is one of the key reasons for moving away from supervised procedures and toward unsupervised ones (Liew et al., 2005). Many bioinformatics techniques, including genome sequencing, rely on counting k-mers (substrings of length k) (Melsted & Pritchard, 2011). k-mers is used to identify the presence of repeated sequences. In k-mer, we either have exact match or no match. Once we divide the protein sequence into k-mers, we need to and compare samples against each other. One technique of comparing this similarity is using the Jaccard similarity. A Jaccard distance of 1 indicates that the samples are identical, whereas a distance of 0 indicates that they are far away from similarity (Baharav et al., 2020). Bloom filter is a statistical data structure that can be used to facilitate dynamic set membership queries that have false positives. It never returns false negative. It detects all k-mers that appear twice or more in a data collection in an exceedingly compact manner, with a low risk of false positives. Bloom filters are commonly utilized in computing applications, although they have just recently been applied in bioinformatics (Melsted & Pritchard, 2011). One recently developed tool for analyzing the DNA and protein sequences in mathematical form is called MATHFEATURE. This tool is recommended by the department of Bio Informatics to be used for this thesis. This tool has different mathematical features to analyze the protein sequence. But the results obtained using the math feature tool are complex to handle at this stage. This tool can be further explored in Thesis II [1].

## 

## MOTIVATION

The application of computer-based techniques to the understanding of biological processes is what bioinformatics is all about. Today's technology allows for high-throughput gene and protein analysis, necessitating the use of informatics to solve biological challenges. It has become a major industry emphasis, especially in the post-genomic era. In biological research, a computer-based technique is now commonplace. This research will greatly help to understand the protein sequences by applying NLP and ML techniques

## SIGNIFICANCE

This study has medical significance. The given bacteria is a pathogen. Pathogen spreads disease. All the variants of this pathogen are not equally dangerous. Some are more dangerous; some are less dangerous and some might be neutral. Once the signature of each strain is identified, our study will open further research work to medically diagnose the severity of the strains.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 MATHFEATURE: FEATURE EXTRACTION PACKAGE FOR BIOLOGICAL SEQUENCES BASED ON MATHEMATICAL DESCRIPTORS

(Bonidia et al., 2020) Machine learning techniques have been used to pull out new and relevant knowledge from biological sequences with great success. However, how the sequences are represented has a big impact on how well these algorithms anticipate. As a result, the most difficult problem is figuring out how to quantitatively describe a biological sequence in a numerical vector form using well organized mathematical expression. For biological sequences, different feature extraction strategies have been developed, with the majority of them being available in feature extraction tools. However, there are useful techniques based on mathematical descriptors, such as Fourier, entropy, and graphs, that are not available in present packages. As a result, this study introduces Math Feature, a novel package that includes mathematical descriptors capable of extracting meaningful information from biological sequences. Math Feature offers 37 different mathematical descriptors among these 20 descriptors are grouped into five domains. Numerical mapping, Chaos game, FT, Entropy and Graphs. Math Feature also has descriptors for protein sequences, which predict structural features along the amino acid main sequence. Math Feature is, to our knowledge, the first package to give such a large and complete range of feature extraction techniques for DNA, RNA, and Proteins based on mathematical descriptors

Different mathematical descriptors are used for testing purpose. The results obtained are complex to understand at this level. This tool needs to be further explored in thesis II. Some examples are shown in the experimental section. The details of the Math Feature categories and descriptors is given in the table section.

## 2.2 BASIC LOCAL ALIGNMENT SEARCH TOOL

(Altschul et al., 1990) Basic local alignment search tool (BLAST), is a new perspective for quick sequence comparison. The core technique is straightforward and it may be implemented in a variety of ways and used in a variety of situations, including simple DNA and protein sequence database searches. Existing sequence comparison techniques and tools are slower than BLAST. This tool is widely used in Bio Informatics industry professionals for sequence comparison.

For our research thesis, this tool is of limited importance because we have to find the pattern in the given protein sequence but this tool does not provide this facility. It can be used to compare two sequences and on the basis of the similarity segment pair score, it will show the result. This tool has other versions also such as PSI-BLAST and PHI-BLAST. The procedure of BLAST is shown in the experimental section

## 2.3 EFFICIENT COUNTING OF K-MERS IN DNA SEQUENCES USING A BLOOM FILTER

(Melsted & Pritchard, 2011) Many bioinformatics algorithms use the counting of k-mers (substrings of length k). The memory capacity of conventional computers can easily be exceeded while counting the k-mers in big contemporary sequencing data sets. In today's data files, a significant portion of the space is used to store k-mers that are often only seen once in the data. For many algorithms, these singleton k-mers are useless.

The Bloom filter is a probabilistic data structure that can be used to handle dynamic set membership queries that contain false positives. It helps to detect all k-mers that appear more than once in a data collection in an exceedingly compact manner, with a very low risk of false positives. Bloom filter has been widely used in the industry but very rarely used in Bio Informatics

Bloom filter is used to count non-unique k-mers and a hash table is required to store them. Bloom filter functions as a "staging area" for the k-mers we have seen so far, while the hash table stores all the k-mers that have been seen at least twice. The goal is to save all k-mers seen thus far implicitly using the memory-efficient Bloom filter, while only entering non-unique k-mers into the hash table. An example of Bloom filter has been shown in the experimental section.

## 2.4 SPECTRAL JACCARD SIMILARITY: A NEW APPROACH TO ESTIMATING PAIRWISE SEQUENCE ALIGNMENTS

(Baharav et al., 2020) In genomic sequence processing, alignment of sequences in pairs is frequently a computational barrier, especially in the context of third-generation sequencing technology. The pairwise k-mer Jaccard similarity is occasionally used as a surrogate for alignment size to filter pairs of reads, and min-hashes are used to efficiently estimate these similarities to speed up this process. Jaccard similarity is no longer a reasonable solution for alignment size when the k-mer distribution of a dataset is highly non-uniform. Spectral Jaccard Similarity, a min-hash-based approach for predicting alignment sizes that naturally accommodates for unequal k-mer distributions is used. Min-Hash is a technique that is used to quickly estimate the similarity between two sets.

## 2.5 A REVIEW OF FEATURE SELECTION TECHNIQUES IN BIOINFORMATICS

(Saeys et al., 2007) In bioinformatics, feature selection (FS) approaches have advanced from being an instructive example to being a practical requirement. Constructing a model, the high-dimensional character of the data, in particular, there are numerous modelling jobs in bioinformatics, ranging from sequence analysis to data mining. From microarray analysis to spectral analyses and beyond, A plethora of feature selections has resulted from literary mining. In the field, there are a variety of techniques that are being given.

We will concentrate on the use of features in this review. Techniques based on projection, for example, can be used to reduce the size of a dataset. Compression (e.g., using principal component analysis). Feature selection approaches do not change the original variables, but only choose one. a portion of them. Feature selection in unsupervised learning is very complex topic and the research community is working to find the different techniques.

Feature selection strategies can be divided into three categories based on how they combine feature selection with the building of the model: filter procedure, wrapper procedure and embedding procedure. Bioinformatics has a long history with sequence analysis. Two sorts of challenges can be recognized in feature selection: content investigation and signal investigation. Content analysis examines a sequence's general properties, such as its proclivity for coding for proteins or its ability to perform a certain biological function. The goal of signal analysis is to find significant themes in the sequence, such as gene structural elements or regulatory areas. Different software packages are also available for the feature selection such as WEKA, GALGO, PCP.

## GAP ANALYSIS

# The widely used tool for the comparison of the protein and DNA sequences is the BLAST that is available online. As presented in the experimental section, BLAST tool can be used to search for a single query. For example, one protein sequence can be given to the BLAST tool and the tool will return all the sequences that matches the query from the database. This is a very basic tool for understanding purpose. BLAST does not provide any mechanism for pattern matching. For our research thesis, we need a tool that take thousands of sequences as input and find the similar pattern among those sequences.

## PROPOSED METHODOLOGY

The methodology for the thesis is still in the finalization stage. As the required pattern can be anywhere in the protein sequence, NLP techniques are required to dig out the pattern in each strain and then Machine Learning algorithm can be applied for the classification

## BLOCK DIAGRAM

MATH FEATURE

STRAIN SIGNATURE

FEATURE EXTRACTION

PROTEIN SEQUENCE DATASET

NLP TECHNIQUES

CLASSIFICATION

UNSUPERVISED ML ALGO

# EXPERIMENTAL SECTION

## MATH FEATURE

This tool can be installed in both Windows and Ubuntu. For testing purpose, we have used Ubuntu. First download the math feature tool and unzip it.

Enter the below command into the terminal. System will automatically show the packages that needs to be installed.

cd MathFeature

cd GUI

cd python3 main.py

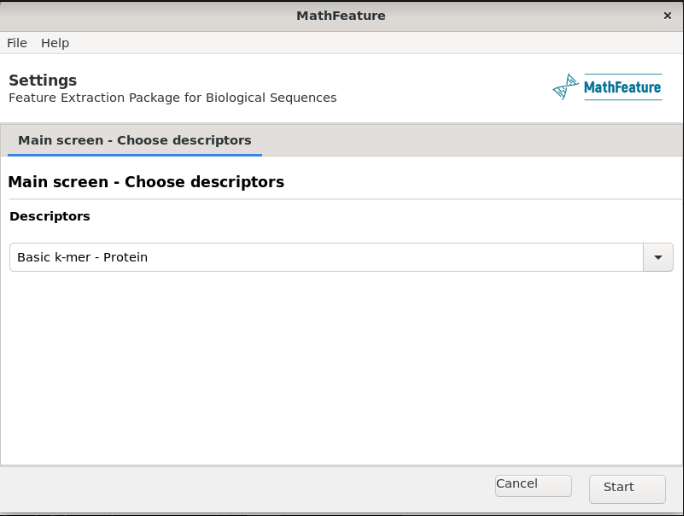
pip install gooey (wxpython will also be installed)

pip install Bio

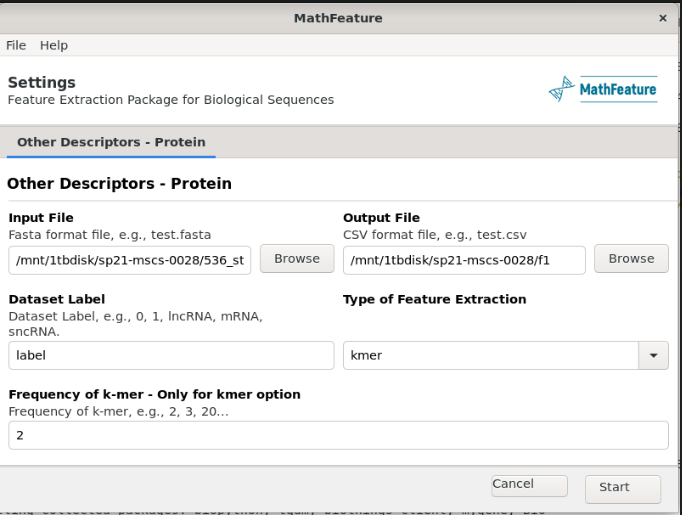
once installation is done, type below command to start Math Feature

python3 main.py

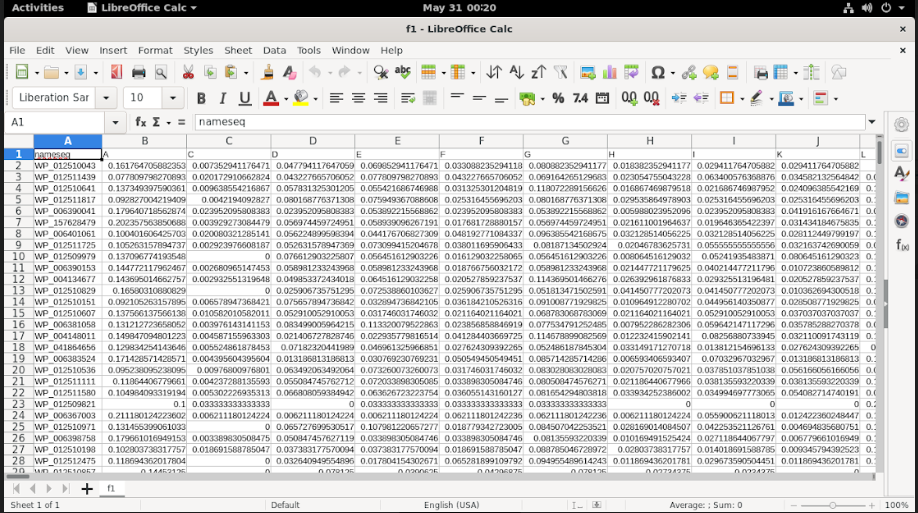
Math Feature GUI will appear. Select the descriptor as required. For testing purpose, we have selected Basic k-mer - Protein



Enter all the required details and press start



Once the process is completed, output file will be created as shown below



As we had selected the k-mer frequency as 2, the system will generate all the single k-mer results and in the combination of two. This file shows the occurrence of each available amino acid in the sequence.

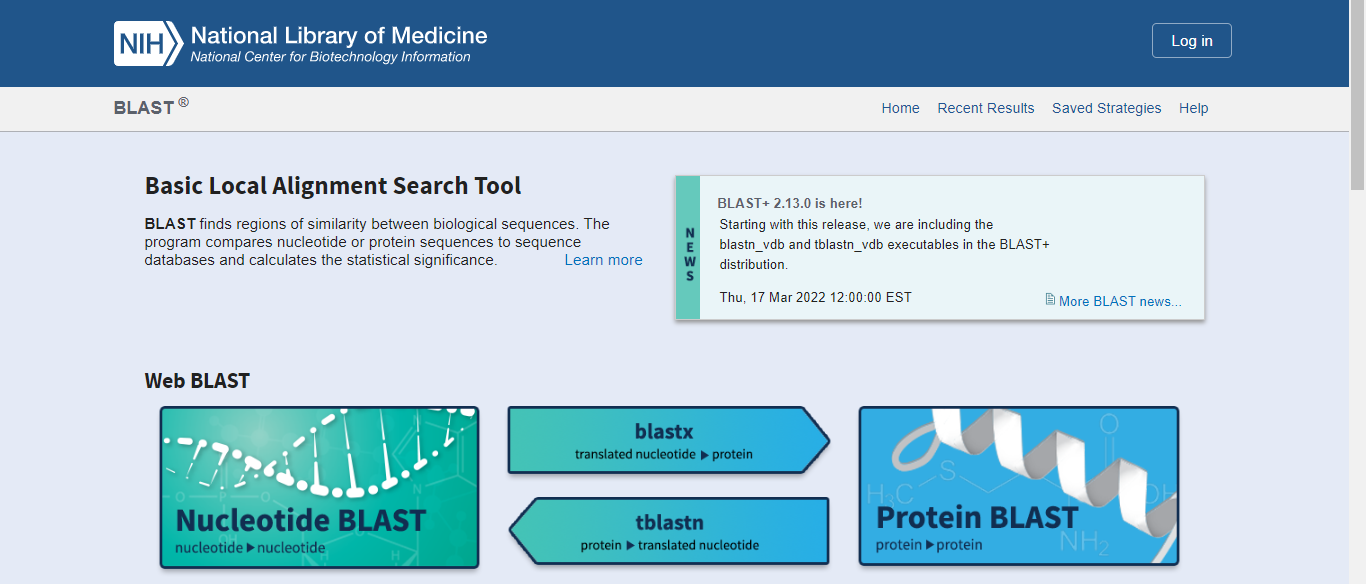
## BASIC LOCAL ALIGNMENT SEARCH TOOL

As discussed in the literature review, it is an online tool that is widely used by the industry experts for DNA / Protein sequence comparison. For the testing purpose, we have given one sample protein sequence to the BLAST to analysis the obtained results

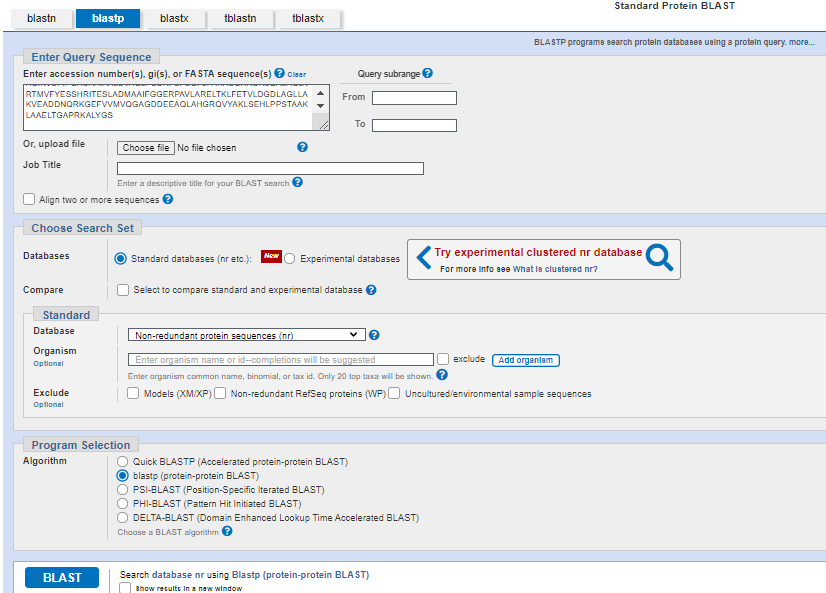
Navigate to the following URL

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

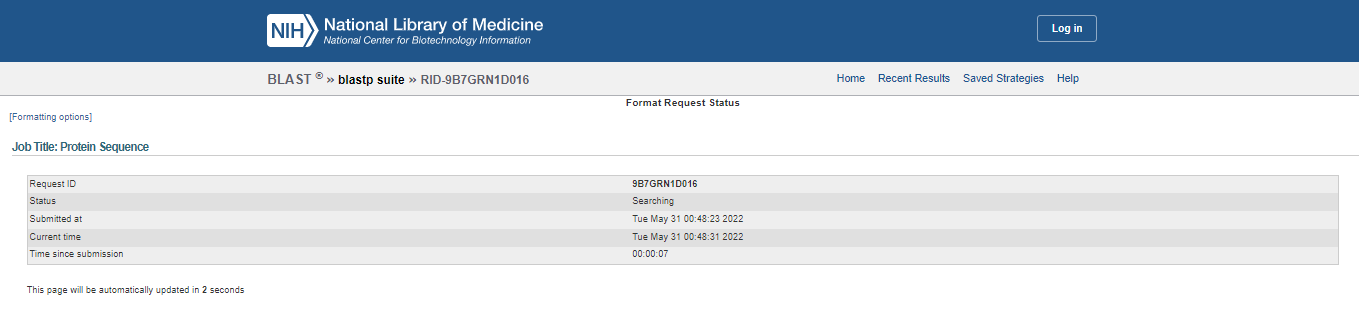
Select Protein BLAST



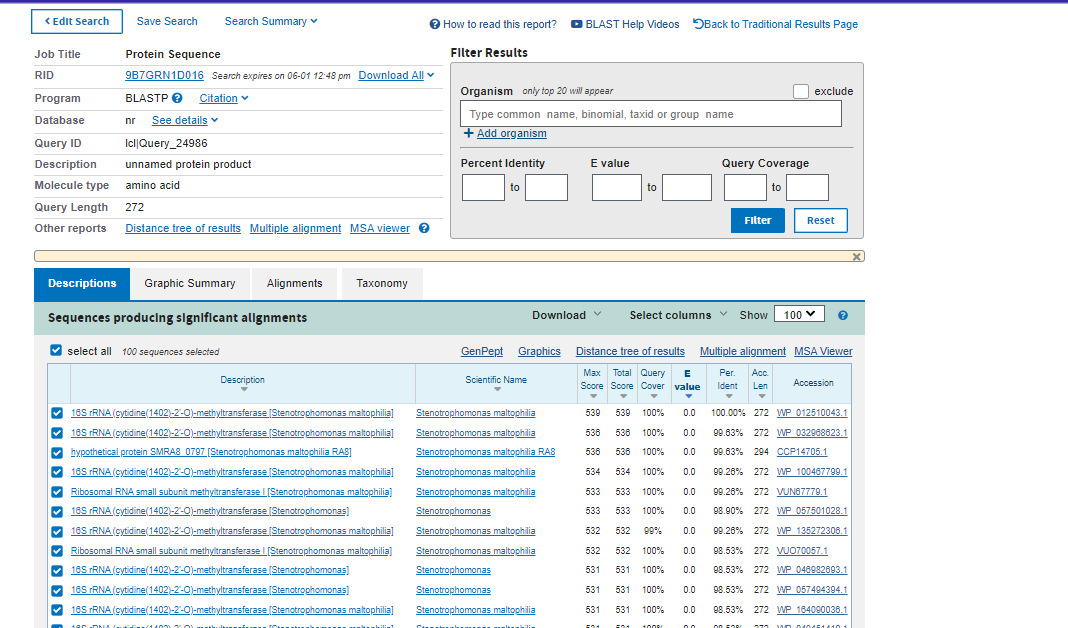
In the Query Sequence, enter the protein sequence and in the Algorithm select blastp and press BLAST button



The system will search this protein sequence in the database



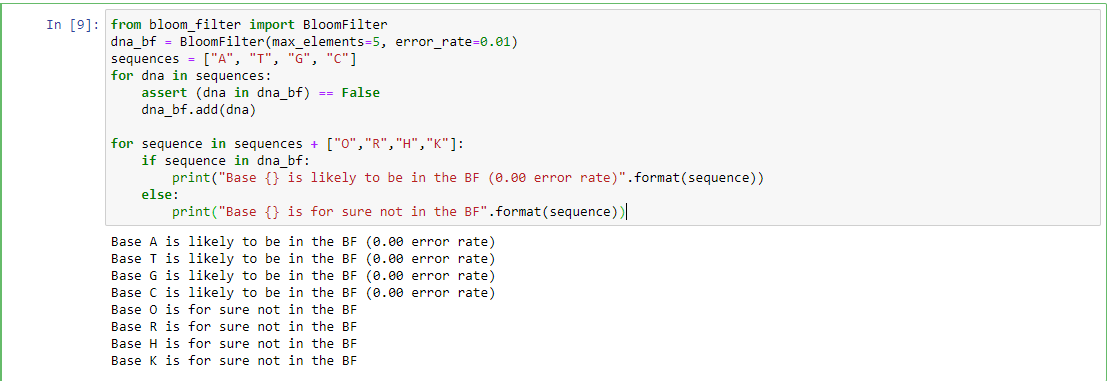
Once the process is completed, the system will show the result based on the score



The result gives details like molecule type, query length, scientific name of the protein, Max score, Total Score, Query cover. We can select other variants of BLAST such as PSI-BLAST, PHI-BLAST. But the gap with the usage of this tool is that it can only compare the given sequence with the available sequences in the database. It cannot identify any pattern in the sequence.

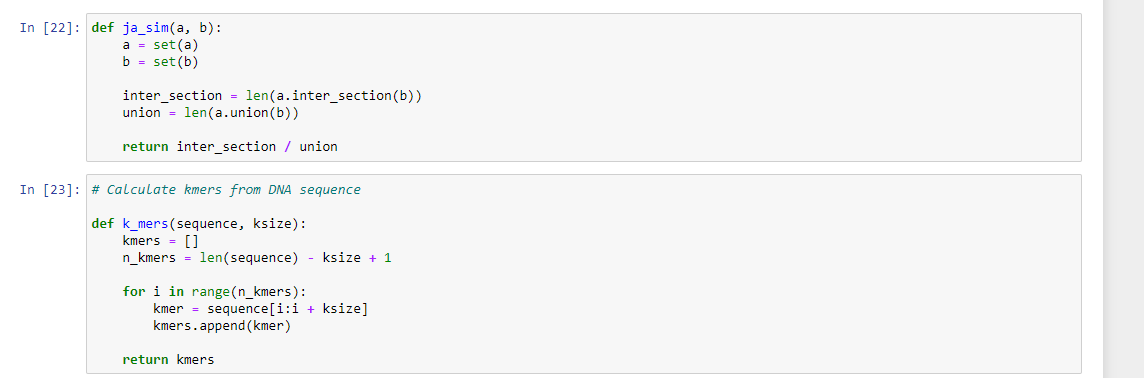
## BLOOM FILTER

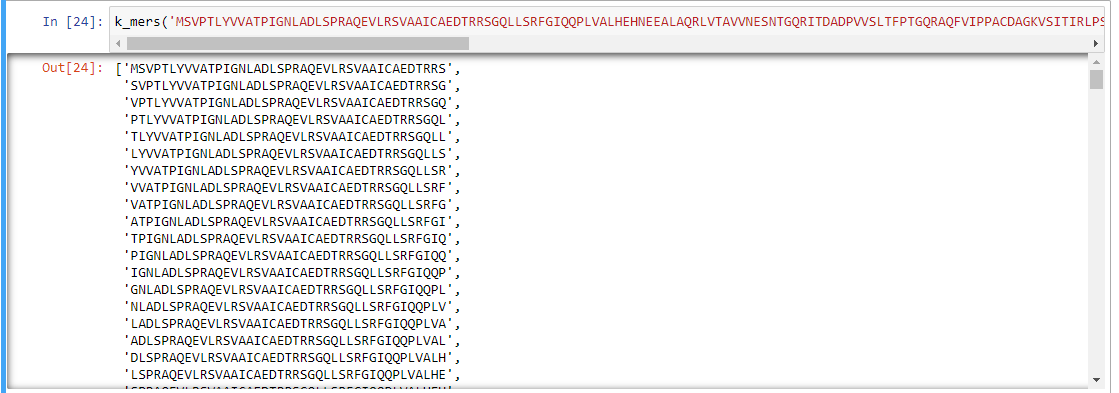
As the paper explains, this approach has been used in the industry but very rarely in the Bio Informatics domain. As an example, we have given four DNA elements to the query and then analyze how this algorithm works

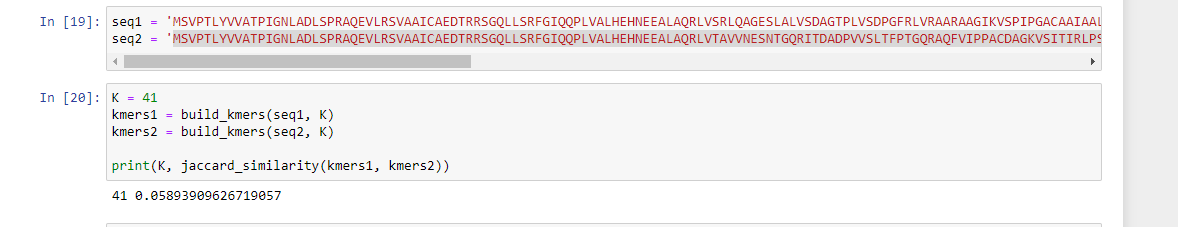


## K-MERS FOR GENOME COMPARISON USING JACCARD

Jaccard similarity can be used to compare two sequences. First the sequences need to be broken into k-mers, then these k-mers can be compared using Jaccard Similarity







# References

1. Bonidia, Robson P., Danilo S. Sanches, and André CPLF de Carvalho. "Mathfeature: feature extraction package for biological sequences based on mathematical descriptors." *bioRxiv* (2020).
2. Ofer, Dan, Nadav Brandes, and Michal Linial. "The language of proteins: NLP, machine learning & protein sequences." *Computational and Structural Biotechnology Journal* 19 (2021): 1750-1758.
3. Liew, Alan Wee-Chung, Hong Yan, and Mengsu Yang. "Pattern recognition techniques for the emerging field of bioinformatics: A review." *Pattern Recognition* 38.11 (2005): 2055-2073.
4. de Ridder, Dick, Jeroen de Ridder, and Marcel JT Reinders. "Pattern recognition in bioinformatics." *Briefings in bioinformatics* 14.5 (2013): 633-647.
5. Saeys, Yvan, Inaki Inza, and Pedro Larranaga. "A review of feature selection techniques in bioinformatics." *bioinformatics* 23.19 (2007): 2507-2517.
6. Altschul, Stephen F., et al. "Basic local alignment search tool." *Journal of molecular biology* 215.3 (1990): 403-410.
7. George, Anon, et al. "Significance of global vectors representation in protein sequences analysis." *Computer Aided Intervention and Diagnostics in Clinical and Medical Images*. Springer, Cham, 2019. 261-269.
8. Melsted, Pall, and Jonathan K. Pritchard. "Efficient counting of k-mers in DNA sequences using a bloom filter." *BMC bioinformatics* 12.1 (2011): 1-7.
9. Cheng, Lok-Lam, David W. Cheung, and Siu-Ming Yiu. "Approximate string matching in DNA sequences." *Eighth International Conference on Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings.*. IEEE, 2003.
10. Kocaleva, Mirjana, et al. "Pattern recognition and natural language processing: State of the art." *Tem Journal* 5.2 (2016): 236-240.
11. Baharav, Tavor Z., et al. "Spectral Jaccard Similarity: A new approach to estimating pairwise sequence alignments." *Patterns* 1.6 (2020): 100081.

# PLAGIARISM REPORT