|  |  |  |
| --- | --- | --- |
| **Research Title *(English)*** | **Application of Discrete Mathematics in DNA Sequence Compression** | |
| **Team** | Marina Reda Abdullah Mekhael (221101235) Mohamed Ahmed Fathi Amer (221101699) Hamza Nashaat Abdelbaki Abdelnabi (221100328)  Aly Maher Abdelfattah Abdelrahman (221101789) Mohamed Mosad Fawzy (221100611) | |
| **Field** | Computer science and Engineering | |
| **Program** | Artificial intelligence science | |
| **Instructor** | **Dr. Mohamed Abd Elaziz** | |
| **This part for the instructor: Notice** | | **Degree** |
| **Date of Submission** |  | |

# Introduction:

DNA sequence data is one of the most important and complex types of data in the field of genetics. These sequences contain genetic information that helps in understanding the evolution of different species and diagnosing genetic diseases. However, with the increase in the volume of DNA sequence data, efficient storage and transmission of this data have become a challenge. In this proposal, we aim to explore the application of discrete mathematics in DNA sequence compression.

DNA is a complex biological molecule consisting of four different nucleotide bases, commonly denoted by the letters A, C, G, and T. These nucleotides can be arranged in a specific order to encode genetic information. Due to the vast amount of genetic information contained in DNA, it is often necessary to compress it for storage and analysis.

The compression algorithm implemented in this project uses Discrete Mathematics concepts to achieve compression. Discrete Mathematics is a branch of mathematics that deals with discrete objects, which are distinct and separate. In this case, the nucleotide bases are considered as discrete objects.

The algorithm first converts the nucleotide bases into binary code using a dictionary, where A is represented by 00, C by 01, G by 10, and T by 11. It then applies various compression techniques, including run-length encoding and Huffman coding, to reduce the size of the binary code.

Run-length encoding is a simple compression technique that involves replacing a repeated sequence of characters with a count and the

character. For example, the sequence “AAAAACCCCCGGGGGTTTTT” can be compressed to “5A 5C 5G 5T”. This reduces the size of the code by reducing the number of characters needed to represent it.

Huffman coding is a more complex compression technique that involves assigning shorter codes to more frequent characters and longer codes to less frequent characters. This reduces the overall length of the code by using shorter codes for the most common characters.

The implementation of this DNA compression algorithm using Discrete Mathematics techniques shows the practical application of mathematical concepts in solving real-world problems.

# Objectives:

* To analyze the characteristics of DNA sequences and identify the patterns that can be exploited for compression.
* To design and implement a compression algorithm based on the identified patterns and using concepts from Discrete Mathematics.
* To evaluate the performance of the developed algorithm and compare it with existing DNA compression techniques.

# Methodology:

* Data collection: DNA sequences will be collected from public databases such as NCBI.
* Data preprocessing: The collected DNA sequences will be preprocessed to remove any non-DNA characters and convert the sequences to a suitable format for compression.
* Pattern analysis: The preprocessed DNA sequences will be analyzed to identify the patterns that can be exploited for compression. This analysis will involve the application of concepts from Discrete Mathematics such as graph theory, combinatorics, and information theory.
* Compression algorithm design: Based on the identified patterns, a compression algorithm will be designed using concepts from Discrete Mathematics such as Huffman coding, arithmetic coding, and run-length encoding.
* Performance evaluation: The developed algorithm will be evaluated using standard metrics such as compression ratio, compression speed, and decompression speed. The performance of the developed algorithm will also be compared with existing DNA compression techniques.

# Conclusion:

The proposed project aims to develop an efficient compression technique for DNA sequences using concepts from Discrete Mathematics. The project will involve the analysis of DNA sequences to identify the patterns that can be exploited for compression, and the design and implementation of a compression algorithm based on the identified patterns. The project is expected to lead to the identification of new patterns in DNA sequences and the development of a more efficient DNA compression technique.