# Pattern Matching Compression Algorithm for DNA Sequences

## Assignment 2: Related Works

The huge size of DNA databases consisting of complex and logical structures results in the need for a large storage space. This can be dealt with a powerful compression algorithm. The DNA is split into blocks and compressed based on the occurrence of the patterns. The ASCII encoding is done where the index is taken into account that compares one of the text characters with a specified position with the pattern character, then based on the matching numbers, the frequency is calculated and the ASCII encoding occurs. In compression technology, mathematical algorithms are used, and compression is achieved without loss of data through (replacing the sub-strings of the repetition sequence from the dictionary, the duplicates are identified through algorithms by looking at the previous sequences). LZ-based indexing is one of the effective indexing techniques that helps eliminate duplication. A combination of reference-based and clustering algorithms (De novo clustering) was used and the (Quip) algorithm was used due to the fact that the amount of space consumed by the memory is very less. The CABLASTP algorithm developed a database that could store data

that was formed by combining two algorithms (dictionary-based compression where ASCII substitutions, reduced repetitive protein sequences, and sequence alignment). There are two main applications for compression of genomic data, improving the performance of De novo clustering in constructing genome segments from a first sequence. Debruijn diagrams are read without reference to genomes by reducing memory consumption.

It was observed that only two of the compressors (DELIMINATE, MF Compress), and the FASTA format generally used, deal with practical operations such as (compressing, decompressing vertebrate genomes completely.)

Different data sets are used in health care systems, especially the DNA sequence used in different health care processes such as (DNA pattern analysis, patient diet recommendation system).