Pattern Matching Compression Algorithm for DNA Sequences

By/ Samaa Mahmoud Mohammed

Dr/ Sara El-Metwally

Abstract

With the advancement in DNA research, it has led to many difficulties in the process of data transfer, maintenance and storage, and this requires an abundant storage space that needs to be dealt with and reduced with effective methodological assistance . Data compression is one of the techniques that can reduce the size of the DNA sequence and save space. The improved compression algorithm helps in determining the pressure of the DNA with matching the pattern of the biological sequence and storing the matching information in a dictionary. Therefore, this algorithm leads to a better average compression ratio of 89% .

Introduction

The volume of biological data grows. The field of molecular biology is facing such a challenge as the regular implementation of amino acid or nucleotide sequences for estimation of biological molecules occurs. Nucleic acid consists of 4 nucleotides, which are (Adenine, Guanine, Thymine, and Cytosine). The human genome consists of approximately 3.1647 billion pairs of DNA bases. The first step in a DNA-based research is to look for patterns within the bases of a DNA sequence. There must be DNA sequencing compression to achieve reasonable storage. Compression systems without data loss are composed of three general categories which are :

(1)symbol-wise replacement

(2) dictionary-based methods

(3) context-based methods. Among the methodologies that take into account the repetition structures or any type of intrinsic regularities in the DNA sequence is the (Bio Compress, Gene Compress) algorithm to achieve the compression of the DNA sequence.

Related works

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.

The duplicates are identified through algorithms by looking at the previous sequences). .

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.

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:

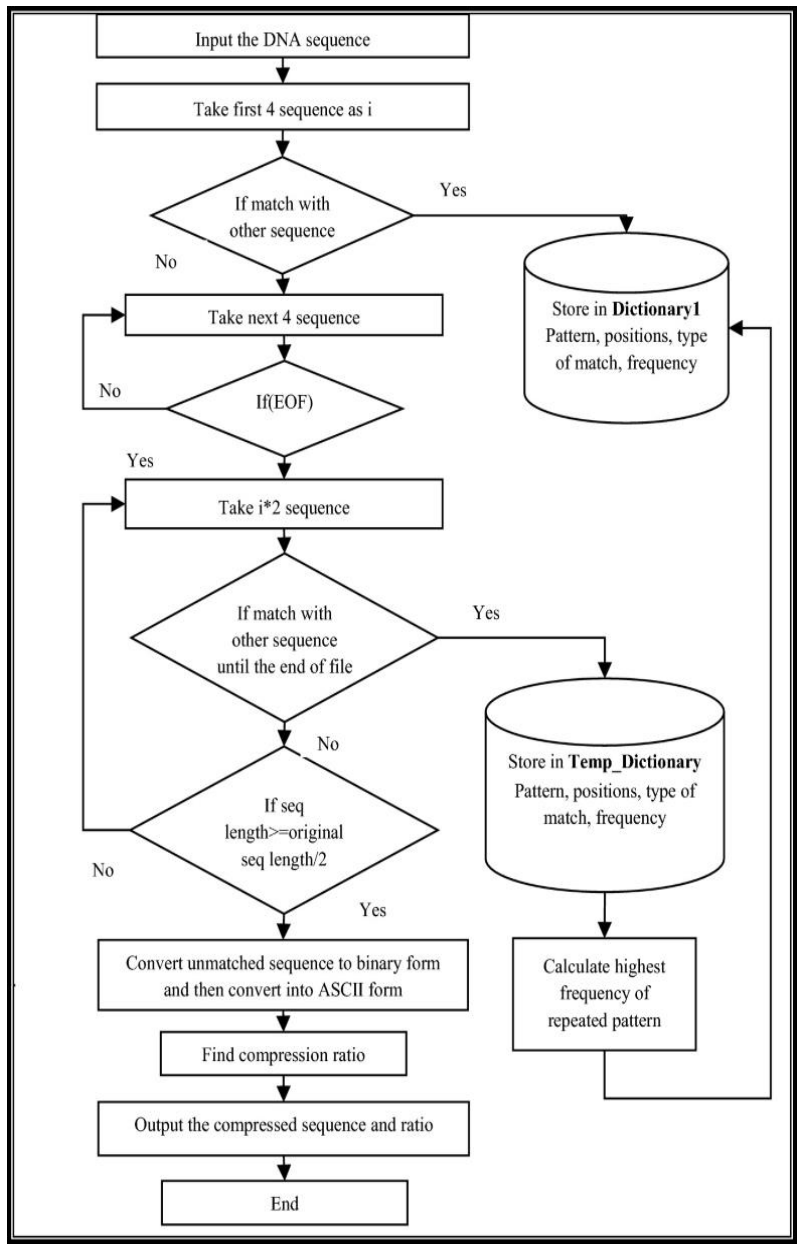
(1)DELIMINATE

(2) MF Compress

(3) FASTA format generally used, deal with practical operations such as (compressing, decompressing vertebrate genomes completely.)

Methodology

The research work includes a text file that comprises four successive base pairs namely '(A,' 'G,' 'T,' and 'C).' The text file helps in performing the matching between the original file and the decompression file. The output file comprises ASCII value which is equivalent to the encoded binary value of the four unmatched base pairs. The architecture diagram of DNA sequence compression is illustrated in Fig. 1. The Dictionary1 (storing highest frequency pattern) .



Results

The Improved Compression algorithm was tested and validated on a group of DNA sequences fetched from NCBI Repository. The standard sequences are Homo sapiens (HUMHDYSTROP), human growth hormone (HUMGHCSA), human beta globin region on chromosome 11(HUMHBB), human DNA sequence (HUMHDABCD), (HUMHPRTB), and vaccinia virus Copenhagen complete genome (VACCG). The testing was carried out on the above data for different sequence lengths via Improved Compression algorithm. The research includes an implementation which is being performed on JAVA environment and the entire dataset that is retrieved from different notepad file is taken as an input and been managed using the MYSQL database for storing and retrieving the results.