Pattern Matching Compression Algorithm for DNA Sequences

BY:Samaa Mahmoud mohammed,Dr:Sara El-metwaly

Abstract

DNA contains genetic information about biological species and DNA databases as a repository for a huge collection of DNA sequences and 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% compared to the ratio. Current dictionary based algorithm compression and ASCII encoded value.

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 symbol-wise replacement, dictionary-based methods and 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 Bioc Compress, Gene Compress algorithm to achieve the compression of the DNA sequence.