Abstract

The genome of an organism contains all hereditary information encoded in Deoxyribonucleic Acid (DNA). Molecular sequence databases (e.g., EMBL, Genbank, DDJB, Entrez, SwissProt, etc) represent millions of DNA sequences filling many thousands of gigabytes and the databases are doubled in size every 6-8 months, which may go to beyond the limit of storage capacity. There are several text compression algorithms used for DNA compression. This paper proposes a new hybrid algorithm is used to compress DNA sequence, the algorithm is designed by combining the fixed length binary code with the LZW (Lempel-Ziv-Welch) compression algorithm. Initially the input sequence is divided into fragments where each fragment consist of four nucleotides and fixed length binary code is assigned to each nucleotide then the pattern (STR and CHR) in LZW used the same for creating the dictionary. Assigning a new binary code for each pattern in the dictionary using a binary tree, and the sequence is replaced binary code for the longest match in the dictionary while compression. The proposed approach attains maximum compression in DNA sequences.

References

- Raja Rajeswari and Dr. Allam Apparao "Genbit compress – algorithm for repetitive and non-repetitive DNA sequences" Journal of Theoretical and Applied Information
A Vital Approach to compress the Size of DNA Sequence using LZW (Lempel-Ziv-Welch) with Fixed Length Binary Code and Tree Structure

Technology 2005


Index Terms

Computer Science

Information Sciences

Keywords

Embl Ddjb Genome Deoxyribonucleic Acid (dna) Lzw Nucleotide