Abstract

One of the most difficult challenges in lossless data compression is finding the right model for the Deoxyribonucleic Acid (DNA) compression. DNA sequences include four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T) where the information in DNA is stored as a code made up of these four chemical bases and these sequences show these are not random, if they are totally random then store them in two bits, this is the most efficient and logical way. This paper proposed an algorithm called A2 for DNA data compression. The proposed algorithm consists of four stages to build a substitutional model. The first stage used a modified version of run-length coding, in second and third stages mapping model for formatting data to be suitable for the final stage fed into Burrows-Wheeler Transform to use permutation technique that group related symbols as possible to improve dictionary coding using Lempel-Ziv (LZ77) and output file stored as (.a2) extension. The A2 algorithm implemented and tested on data from GenBank and shows acceptable file size and processing time ratio.

References
Propose a Substitution Model for DNA Data Compression


Index Terms

Computer Science Information Sciences

Keywords
DNA compression, Compression Model, Run-length coding, Mapping Table, BWT, Dictionary Coding, A2 algorithm.