International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 5 | ISSN : 2456-3307

A Relative Study on Existing Two Bit-Based DNA Compression Techniques with Bit DNA Squeezer (BDNAS)

Alam Jahaan[1], Dr. T.N. Ravi[2]

[1]Research Scholar, Department of Computer Science, PERIYAR EVR College, Trichy, Tamilnadu, India
[2]Assistant Professor, Department of Computer Science, PERIYAR EVR College, Trichy, Tamilnadu, India

ABSTRACT

Compression can ease the burden of storage, transfer, retrieval and searching of data. DNA databases are growing intensively, leading to the demand for more sophisticated compression tools. Compressing DNA data is different from compressing any other text since DNA sequences are essentially made up of combinations of four nucleotides Adenine (A), Thymine (T), Cytosine (C), or Guanine (G) which queue up in a particular sequence and make a long string of repetitive and non- repetitive sequences. Recently, the explosion in Bio Informatics has drawn attention towards DNA storage and banking. DNA banks are growing rapidly as newer samples of DNA sequences are being accumulated frequently. This paper deals with comparative study of existing two bit based DNA compression techniques such as GENBIT, DNABIT, HUFFBIT, GENCODEX with one bit based BDNAS (Bit DNA Squeezer) algorithm.

Keywords : DNA, bases, bit based DNA compression, BDNAS, Compression ratio.

I. INTRODUCTION

Compression is the process of encoding data using a technique or algorithm that shrinks the total size of data. It basically, is of two types Lossy compression where the decompressed image is not exactly as the original one and Lossless compression where the originality of the compressed image after decompression is retained [1]. Lossless compression is suitable for DNA compression because it preserves the originality of the DNA sequence after decompression. The main task involved in compressing DNA sequences is to eliminate redundancy.

1.1 DNA Compression

All living organisms are made up of cells, which contain DNA (Deoxyribonucleic acid) which is necessary for the growth, development and functioning of all known living organisms. The order or sequence, of these bases determines the information available for building and maintaining an organism [2]. DNA Data Banks or DNA databases are used to store DNA sequence and occupy more storage when compared to other non DNA databases[3].The enormous genome datasets needs to be stored in its original form and requires large storage space. Here compression becomes appropriate. As DNA datasets cannot afford to lose any part of their data, Lossless compression techniques are more suitable for their compression. Lossless data compression simply finds the most efficient coding for the data by eliminating redundancies and reproducing the exact copy of the original form on decompression [1]. DNA sequences contain repetitions of A, C, T, G. which are random sequences that contain long-term repetitions with sub sequences similar to each other. Compression of DNA sequences involves searching for repetitions, encoding them and storing or transmitting them. On requirement, the encoded sequences will be decoded to their original form [4]. The properties recognised in most of the sequences are the oft-repeated substrings, repeated palindromes and repeated reverse compliments [5] hence leading to researching and developing newer sophisticated DNA compression techniques. Recently two-bit coding methods have become popular where the

four nucleotide bases {A, C, G, T} in DNA sequences are assigned values 00, 01, 10 and 11 respectively [6].

1.2 Organization of the paper:

This paper compares existing two bit based DNA compression techniques with the one bit based DNA technique BDNAS. Existing two bit based techniques and BDNAS methods are highlighted briefly in section 2. Performance in terms of compression ratio is evaluated and Compression ratio for Best case, Average case, and Worst case of the techniques like GENBIT, DNABIT, HUFFBIT, GENCODEX and DNACRMP with BDNAS are compared in section 3. Results are tabulated in section 4. Conclusion and future work are discussed in section 5.

II. Existing DNA Compression Techniques:

Main criteria for compression of DNA sequences or any other data is to save storage space, speedy transmission and increase time complexity. DNA sequences contain repetitions of A, C, T, G in the form of a long String with repeated substrings.

2.1 Two-Bit Based methods:

These algorithms implement a bit pre-processing stage by assigning four unique two bits (A=00, G=01, C=10, T=11) to each base before the encoding process. In the following algorithms, each stage is similar in the first (bit pre-processing) stage but different in the coding stage [6].

The GENBIT Compress tool [7] by Rajeswari and Apparao is based on a novel concept of searching and encoding exact repeats by assigning binary bits to the nucleotides before coding. Rajeswari & Apparao later on developed the DNABIT Compress tool [8] it assigns binary bits to exact and reverse repeat fragments of *DNA* sequences. HUFFBIT compress was also proposed by Rajeswari et al., for DNA sequences [9] using the concept of Extended Binary Tree method. GENCODEX introduced by Satyanvesh et al., is a two phased algorithm that produces a better compression ratio at a high throughput by using graphical processing units and multi-cores [10]. In the DNACRAMP tool by Prasad & Kumar, after bit pre-processing, the encoding and decoding process is performed with the help of a two-stage index bounded linear array data structure using basic procedural language [11].

2.2 One-Bit based DNA Compression technique (BDNAS)

DNA sequences are mere combinations of the four-nucleotide bases, which are both repetitive and non-repetitive in nature. BDNAS is an improvement over two bit based DNA compression techniques. The frequency of the nucleotides is computed and depending on the value of occurrences, the bases are assigned 0 or 1 for the first and second occurrences respectively. However, the bases with third and fourth frequencies are assigned 0 and 1 but their locations are recorded in a position map.

The Decompression process, considers the position map along with the compressed file and replaces all the 0s in the compressed file according to the positions recorded in the position map to its corresponding (third) base and replaces the 1s at the positions recorded in the position map to its corresponding (fourth) base. Finally, the remaining 0s in the compressed file are converted to the first base and 1s to the second base respectively. [12].

III. Performance Evaluation

The compression ratio [13] is calculated and compared for the best case, average case and worst case. In all algorithms, the worst-case compression ratio is the highest compression ratio. The best-case compression ratio is the lowest compression ratio and the average case is the compression ratio of a typical input or a random input, which is not given by the worst case nor the best-case efficiency.

Compression Ratio = Size of compressed file / Size of original file.

IV. Results

The compression ratio for best case, worst case and average case is computed for GENBIT, DNABIT, HUFFBIT, GENCODEX, DNACRMP and BDNAS algorithms. Average compression ratios are tabulated in table 1 and Chart 1 depicts the comparisons for compression ratios of the algorithms.

Table1: Comparison of compression ratio

Algorithm	Compression Ratio		
	Best case	Average case	Worst case
GENBIT	1.125	1.727	2.238
DNABIT	1.04	1.53	1.58
HUFFBIT	1.006	1.611	2.109
GENCODEX	0.017	1.42	2.25
BDNAS	1.44	1.462	1.48

The compression ratio of the BDNAS Algorithm is seen to be almost equal for all cases because for any Dataset the size remains the same as the number of bases. Here one bit is assigned per base, unlike the two bit based compression methods where two bits are assigned to a base. A file with 2000 bases will grow into a file with 2000 bits, whereas, in 2-bit based methods a file with 2000 bases will double up into 2000 bits.

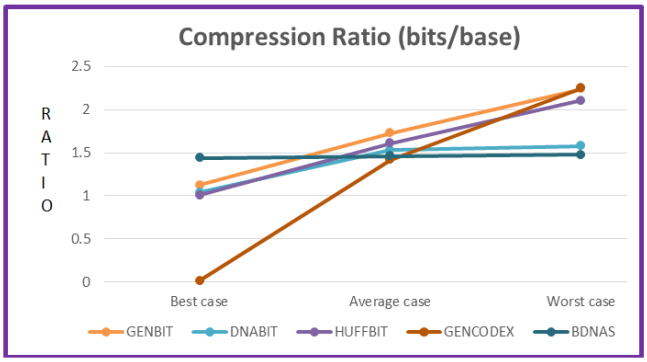

Chart 1: Compression ratio of all three cases

V. CONCLUSION

Owing to the boom in biological and biomedical sciences, along with the demand for storage, transmission and time complexity, DNA compression has become imperative. The average compression ratio for existing two-bit DNA compression algorithms with BDNAS Algorithm has been compared. No standard datasets have been used to calculate the compression ratio. This work may be enhanced by using standard datasets to perform the analysis. More algorithms that are exceptional maybe developed by enhancing BDNAS techniques as an extension for future work.

VI. REFERENCES

[1]. Alam Jahaan ,Dr T.N. Ravi, "Scrutiny Of Lossless Compression Techniques Using A Few Quality Measures", International Journal Of Advanced Research In Computer Science And Applications Issn 2321- 872x, Volume 4, Issue 3, March 2016.

[2]. https://ghr.nlm.nih.gov/primer/basics/dna

[3]. https://en.wikipedia.org/wiki/DNA_database

[4]. Alam Jahaan ,Dr T.N. Ravi, Dr. S. Panneer Arokiaraj, "A Comparative Study and Survey on Existing DNA Compression Techniques", IJARCS, p-ISSN: 0976-5697, volume 8, No.3, March-April 2017

[5]. Manzini G. and Raster0 M., "A simple and fast DNA compressor, Software: Practice and Experience", MUIR support projects(ALINWEB), vol. 34(14), pp.1397-1411, 2004

[6]. Nour S. Bakr, Amr A. Sharawi,"DNA LosslessCompression Algorithms: Review", American Journal of Bioinformatics Research 2013,3(3):72-81, DOI:10.5923/j.bioinformatics.20130303.04

[7]. Rajeswari, P. R., and Apparao, A., 2010," Genbit Compress Tool (GBC): A Java-Based Tool To Compress DNA Sequences and Compute Compression Ratio (BITS/BASE) Of Genomes", International Journal of Computer Science and Information Technology, 2(3), 181-191

[8]. Rajeswari, P. R., and Apparao, A., 2011, "DNABIT Compress – Genome compression algorithm", Bioinformation, 5(8), 350-360

[9]. Rajeswari, P. R., Apparao, A., and Kumar, R. K., 2010, "HUFFBIT COMPRESS – Algorithm to compress DNA sequences using extended binary tree", Journal of Theoretical and Applied Information Technology, 13(2), 101-106

[10]. Satyanvesh, D., Balleda, K., Padyana, A., et al., 2012, "GenCodex - A Novel Algorithm for Compressing DNA sequences on Multi-cores and GPUs", Proc. IEEE, 19th International Conf. on High Performance Computing (HiPC), Pune, India, No 37.

[11]. Prasad, V. H., and Kumar, P. V., 2012, "A New Revised DNA Cramp Tool Based Approach of Chopping DNA Repetitive and Non-Repetitive Genome Sequences", International Journal of Computer Science Issues (IJCSI), 9(6), 448-454.

[12]. Alam Jahaan ,Dr T.N. Ravi, , Dr. S. Panneer Arokiaraj, "Bit DNA Squeezer (BDNAS) : A Unique Technique for Dna Compression", International

Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

[13]. S.R. Kodituwakku Et. Al. "Comparison Of Lossless Data Compression Algorithms For Text Data", Indian Journal Of Computer Science And Engineering, Vol 1 No 4 416-425