Volume 8, No. 3, March – April 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

A Comparative Study and Survey on Existing DNA Compression Techniques

Alam Jahaan
Research Scholar in Computer Science,
PERIYAR EVR College
Trichy, Tamilnadu

DR. T.N. Ravi
Assistant Professor in Computer Science,
PERIYAR EVR College
Trichy, Tamilnadu

Dr. S. Panneer Arokiaraj
Associate Professor in Computer Science,
PERIYAR EVR College
Trichy, Tamilnadu

***Abstract*:** Data storage and DNA banking are receiving tremendous attention recently, due to the explosion of genetic research, biomedical research and forensic sciences.DNA databases when compared to other databases occupy more storage due to the enormous size of each DNA sequence. They are growing exponentially as newer samples are being collected and stored. This poses a major task for storage, transfer, searching and retrieval of data. Hence,Compression of DNA Databases have become relevant.DNA sequences contain repetitions of A, C, T, G,so compression of the sequence involves searching and encoding these repetitions,then decoding them while decompression.This paper explores a few existing DNA sequence compression algorithmswhich were developed based on certainprevailing generalcompression algorithms. First a few, existing compression techniques are explained and then, fiveavailable DNA sequence compression algorithms are evaluated for their suitability in compressing large collections of DNA sequences.

***Keywords*:** DNA, DNA databases, Genetics, DNA compression, Biocompress, Cfact, DNAcompress, Gencompress, CTW+LZ.

I. INTRODUCTION

1.1Introduction to Genetics: Genetics is the study of genes. Genes are present inside the nucleus of a cell and are made up of molecules that are wound together in such a way that the sequence carries information. Thisdetermines how living organisms inherit features from their parents governed by the genes.

Genes are made up of DNA (Deoxyribonucleic acid)which is divided into separate pieces called chromosomes. Chromosomes are packed and ordered structures containing most of the DNA of a living creature. They mainly contain four nucleotides, namely.,adenine (A),cytosine(C), thymine (T), guanine(G), which line up in a particular sequence and make a long string. [1]

DNA is a double helix where C on one string is always opposite to G on the other string; A is always opposite to T. The human genome comprises of about 3.2 billion nucleotide pairs on all the human chromosomes. The order of the nucleotides carries genetic information, whose rules are defined by the genetic code, which controls inheritance and gene expression.

1.2 Introduction to compression: Compression is the technique of reducing the data storage size without degrading the quality of the data. It is an act of expressing the given large dataset, using less storage space.This may be achieved by eliminating redundancy of information while the resultant retains the information of the original data [2].

Data compression in general is divided into reversible compression or irreversible compression. The data may be reproduced into its original form as in reversible otherwise called Lossless compression or it may not be reproduced to its original form namely Lossy compression.

1.3 Organization of the paper: This paper sheds light on DNA sequences and compressing these sequences for storage, sharing and transporting effectively and efficiently. The compression must be a lossless technique since the data cannot be lost or altered. The combination of DNA and compression is discussed in Section 2.A few existing compression algorithms and new DNA compression algorithms which were developed using the discussed existing algorithms are highlighted in Section 3. Section 4 discusses the evaluation and results followed by the conclusion.

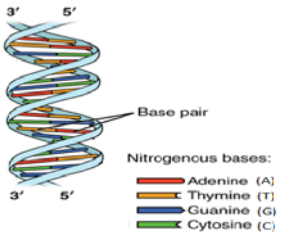

Figure 1: DNA double helix [3]

2.0Essenceof DNA compression: In more recent times, DNA and DNA compression has received marked attention owing to the boom in biological and biomedical sciences.

The essentiality of DNA is important in the field of genetics, genomics, genetic engineering and criminology. DNA not only contains information about ones heritage, it helps to detect the risk for certain diseases and also diagnose genetic disorders.

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 exact copy of the original form on decompression.

DNA Data Banks or DNA databases are used to store DNA sequence and they occupy more storage when compared to

other non DNA databases. DNA databases may be public or private, but the largest ones are national DNA databases[4].

2.1 DNA Compression: Main criteria for compression of DNA sequences or any other data is to save storage space and bandwidth. DNA sequences are random sequences that contain long-term repetitions with sub sequences similar to each other. Compression of DNA sequencescomprises encoding and searching the found repetitions and decoding them when decompressed.

The properties noted in most of the sequences that forms the main criteria for various compression techniques are the oft-repeated substrings, repeated palindromes and repeated reverse compliments[5].

3.0Compression Algorithms

3.1 General Compression Algorithms: Various compression algorithms are used to compress and decompress. The efficiency of any compression algorithm depends on how well and fast it compresses and decompresses, which is generally measured in compression ratio. It may be noted that greater the compression ratio, better the algorithm's efficiency.Some of the several lossless encoding approaches discussed in this paper are:

1. Huffman Encoding
2. Adaptive Huffman Encoding
3. Arithmetic coding
4. Context tree weighted method
5. LZW

Huffman coding, an entropy coding algorithm is used for lossless data compression. It is based on frequency of occurrence of a data item. Pixels of the image or charactersmay beused as symbols. Symbols that occur frequently are assigned a smaller number of bits, while the symbols that occur less frequently are assigned a relatively larger number of bits[2]. Huffman code is a prefix code. The binary code of any symbol is not theprefix of the code of any other symbol, [6].

The Huffman Encoding algorithm is explained with an example below:

- Step 1: Input the string
- Step 2: Sorting the data by frequencies
- Step3:Choose 2 smallest frequencies count.
- Step 4: Merge them together with sum of them and update the data
- Step 5: Repeat step 2, 3, 4.

Adaptive Huffman Coding is expanded on Huffman algorithm that constructs the Huffman tree in one pass, but takes more space than Static Huffman algorithm [7]. Normally, using Huffman coding on standard files shrinks the files from 10% to 30% depending on the character distribution. Huffman encoding creates the tree like structure when it encodes the given string.Coding is examined below.

```
ENCODER:
Initialize_model();
while ((c = getc (input)) != eof)
{
        encode (c, output);
        update_model (c);
}
DECODER:
Initialize_model();
while ((c = decode (input)) != eof)
{
        putc (c, output);
        update_model (c);
}[8]
```

Arithmetic codingmainly defines a method that provides code words with an ideal length. The average code length is close to the possible minimum given by information theory. i.e, it assigns an interval to each symbol whose size reflects the probability for the appearance of this symbol. The code word of a symbol is an arbitrary rational number belonging to the corresponding interval, [9].

Example: The basic idea of arithmetic encoding is to have a probability line, 0-1, After defining the ranges and probability, then encoding starts, every symbol defines where the output floating point number lands.

SYMBOL	PROBABILITY	RANGE
X	2	[0.0, 0.5]
Y	1	[0.5,0.75]
Z	1	[0.75,1.0]

Table 1: Example for Arithmetic coding

The algorithm to compute the output number is:

- Low = 0.
- High = 1.
- Loop. For all the symbols.
  - Range = high - low
    - High = low + range (high_range of the symbol being coded)
    - Low = low + range (low_range of the symbol being coded)
- Range, keeps track of position of next range.
- High and low, specify output number.

The **CTW algorithm** has the context tree which is built dynamically during encoding/decoding process. All of the already visited substrings of shorter size than a fixed bound (the height of the tree), exist as a path in the tree. Each node of the tree contains a probability.Finally the weighted probability is sent to the arithmetic encoder which encodes the symbol, and the encoder goes to the next symbol[10]. Encoder has 2 parts:

- A source modeller (the actual CTW algorithm), which accepts the uncompressed data and estimates the probability on the next symbol.
- An arithmetic encoder, which uses the estimated probabilities to compress the data generated by the actual source.

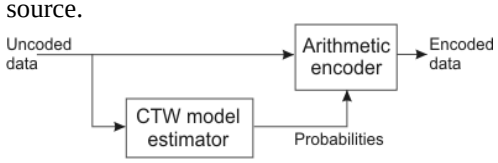

Figure 2: CTW encoder

LZW is invented by Terry Welch in 1984, Initial dictionary contains all alphabets, [11]Algorithm:

```
    p = null
while (!done)
read next symbol int
  if ( p * a) is in dictionary//'*'stands for concatenation
          p = p * a
  else send out index of p
  add p*a to the dictionary
          p = a
end
```

Example:Input: abba_wabba_wabba_wabba_woo_woo_woo

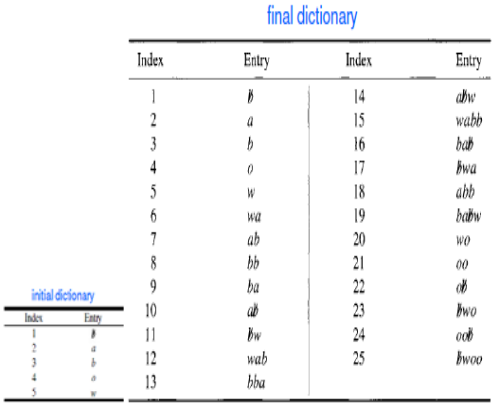

Figure 3: Algorithm &Examplefor LZW

3.1DNA compressionAlgorithms : The algorithms mainly focus on storage and access of the repetitive DNA sequences produced by large-scale genome sequencing projects.

DNA sequences are naturally represented as strings of the four nucleotides A, C, T,G. However, a collection of general-purpose compressors, normally used for text compression, such as gzip orbzip2fail to detect and eliminateDNA-specific redundancy and ultimately result in representations thatrequire more than 2 bits per character. Therefore, tailored DNA compressionmethods are needed and several methods have been proposed to date [12].

Most of the earlier DNA compression methods were concerned with the redundancy within a given DNA string andwere essentially dictionary-based methods or statistical methods.Few of the compression algorithms listed below use one of the above existing encoding approaches to compress and decompress DNA database:

1. BioCompress
2. Cfact
3. GenCompress
4. DNACompress
5. CTW+LZ

Biocompress:One of the first methods proposed for compressing DNA was Biocompress by Grumbach and Tahi (1993) [13], which detects exact repeats and palindromes of arbitrary distance to each other and replaces them with pointers to previous occurrences. Two lossless compression algorithms for DNA sequences, namely BioCompressand BioCompress-2 were primarily created, making use of the Ziv and Lempel data compression method [14]. BioCompress- 2 finds both the exact and reverse repeats in the target sequence. It encodes them by repeat length and the position of a previous repeat occurrence. If there is no significant repetition then the arithmetic coding of order- 2 is used to reduce the number of bits used. The only difference between BioCompress and BioCompress-2 is the use of arithmetic coding.

Cfact [15] is similar to Biocompress and uses a two-pass algorithm to search for the longest exact and reverse complement repeats. It builds the suffix tree of the sequence in the first pass and does the actual encoding using LZ in the second pass. Non-repeat regions are also encoded by 2 bpb (bits per base). The results may be similar to Biocompress, but they take more compression time, since Cfact uses two passes, and constructs a suffix tree.

Gencompress [16] is a one-pass algorithm that searches for the approximate matches. This algorithm uses order- 2 arithmetic encoding. It detects the approximate complemented palindrome (A replaced by T and C replaced by G) in DNA sequences. The average compression ratio is 1.7428 bits/bytes. Gencompress achieves higher compression ratios compared to Biocompress or Biocompress-2.

DNACompress [17] uses Lempel-Ziv compression scheme. It finds all the approximate repeats including complemented palindromes and encodes approximate repeat regions and non-repeat regions. The average compression ratio is 1.7254 bits/bytes.

CTW-LZ (Context-Tree Weighting) algorithm produces a ratio thatoutperforms the Lempel-Ziv (LZ) type of methods. The CTW algorithm actually performs a sophisticated model and parameter estimation, uses a weighting of multiple models to determine the next symbol probabilities. The algorithm detects approximate repeats using dynamic programming thenencodes long exact and approximate repeats using an LZ77-type encoding[18]. The Context-Tree Weighting achieves a very good compression ratio but has a much slower execution speed than e.g. Lempel-Ziv type of algorithms.

3.3 Quantifying compression: The parametric measures of quality are basically used to measure Compression Performances hence they may be called Measurement Parameters which may differ based on the compressed file [19]. To quantify the efficiency of a given compression run, several measures can be applied such as

- Compression Ratio
- Compression Factor
- Data savings
- Compression Time
- Decompression Time

For example, if a 10 MB file is compressed to 2 MB, then the compression ratio is 0.2, the compression factor 5 and the data savings 80%. In order to simplify the test platform we have used only 5 human genome subsets of varying sizes for compression with the above discussed algorithms.

3.3.1 Compression Ratio: The ratio between the compressed file size and the original file size.

$$\text{Compression ratio} = \frac{\text{Compressed size}}{\text{Uncompressed size}}$$

3.3.2Compression Factor: The ratio between the original file and the compressed file. This is essentially the inverse of the Compression Ratio

$$\text{Compressionfactor} = (\text{compression ratio})^{-1}$$

$$= \frac{\text{Uncompressed size}}{\text{Compressed size}}$$

3.3.3Data savings or Savings percentage: It is the percentage of the size reduction of the file, after the compression

Data savings = (1- compression ratio)*100

3.3.4Compression Time: The time taken by the algorithm to compress the file it is calculated in milliseconds (ms).

3.3.5Decompression Time: The time taken by the algorithm to decompress and retrieve the original file from compressed file. It is calculated in milliseconds[20].

4.0 Evaluation and Results of DNA compression algorithms:

4.1 Evaluation : The comparison of compression ratio for above algorithms has been tabulated in table1 and depicted in the graph below for five different Human genome datasets.

Table1:Comparison of compression ratio

DNA SEQUENCE	SIZE BITS	BIOCOMPRESS	GENCOMPRESS	CTW-LZ	DNACOMPRESS
HUMDYSTROP	33770	1.9262	1.9231	1.9175	1.9116
HUMHDABCD	58864	1.8770	1.8192	1.8218	1.7951
HUMHPRTB	56737	1.9066	1.8466	1.8433	1.8165
HUMGHCSA	66495	1.3074	1.0969	1.0972	1.0272
HUMHBB	73308	1.8800	1.8204	1.8082	1.7897

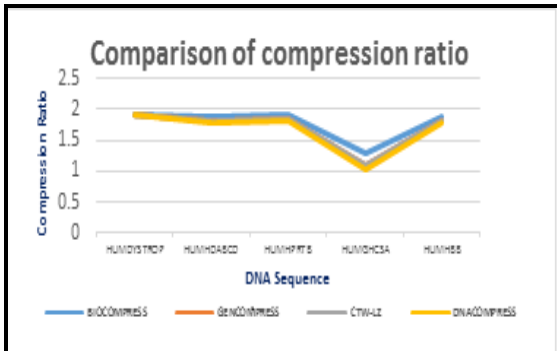

Chart 1: Compression ratio

4.2Results:DNACompress, CTW-LZ and GenCompressobtain the best results among thefiveexisting algorithms. Biocompress has a greater compression compared to the other algorithms. Both GenCompress and DNACompress use the greedy approachfor selection of the repeat segments. There are no compression results ofCfact algorithm; therefore it is difficult to compare [21]. GenCompress selects the best prefixof the region which is not yet encoded to be coded at the next step. CTW+LZ tries to solve the problem of GenCompressbut is time consuming.

CONCLUSION

The above paper elucidates the importance of DNA datasets. Its storage and transfer is crucial thus leading to the need for compression.Existing compression techniques may be used individuallyor blended together to enhance and create new algorithms with better performance in terms of speed and storage space.All the quality measures discussed may be calculated and compared as an extension of future works as only the compression ratio for all the five datasets are tabulated and evaluated.

REFERENCES

[1] https://en.wikipedia.org/wiki/Introduction_to_genetics
[2] AlamJahaan ,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.
[3]http://cnx.org/contents/GFy_h8cu@9.87:U7tPDRxK@7/DNA-Structure-and-Sequencing#fig-ch14_02_03
[4]https://en.wikipedia.org/wiki/DNA_database
[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]. Chang C.J, “Recent Development Of Images Compression Technique”
[7] Singh A. And Gahlawa M. “Image Compression and Its Various” International Journal Of Advanced Research In Computer Science And Software Engineering. Volume 3, Issue 6, June 2013
[8]https://users.cs.cf.ac.uk/Dave.Marshall/Multimedia/node212.htm
[9] Kaur M. And Kaur G. “A Survey Of Lossless And Lossy Image Compression Technique” International Journal Of Advanced Research In Computer Science And Software Engineering Volume 3, Issue 2, Feb 2013
[10] BehshadBehzadi and Fabrice Le Fessant LIX “DNA Compression Challenge Revisited”, EcolePolytechnique, Palaiseaucedex 91128, France .
[11]Mark Nelson, Jean-Loup Gailly, “The Data Compression Book”, Data Compression Techniques University Of Bahrain.
[12] D Satyanvesh, KaliudayBalleda, AjithPadyana, P.K. Baruah “GenCodex - A Novel Algorithm for Compressing DNA sequences on Multi-cores and GPUs”, Sri SathyaSai Institute of Higher Learning, PrasanthiNilayam, India
[13] A Grumbach& F Tahi. “Compression of dna sequences”. In Proceedings of the IEEE Data Compression Conference, Snowbird, UT, USA., March 30April 2. 1993.
[14] J. Ziv and A. Lempel. “A universal algorithm for sequential data compression”. In IEEE Trans. Inform. Theory, volume 23, pages 337–343, 1977.
[15] E. Rivals, M. Dauchet, J-PDelahaye, et al., "Fast DiscerningRepeats in DNA Sequences with a Compression Algorithm,"The 8th Workshop on Genome and Informatics, 1997, vol. 8, 215-26.
[16]Ming Li Xin Chen, Sam Kwong. A compression algorithm for dna sequences. In Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan, April 8-11, 2000
[17] Bin Ma Xin Chen 1, Ming Li and John Tromp. Dnacompress: fast and effective dna sequence compression. volume 18, pages 1696–1698, 2002.
[18] Mr Deepak Harbola, Dr. R.K. Bharti, “State of the art: DNA Compression Algorithms”International Journal of Advanced Research in Computer Science and Software Engineering Research, ISSN: 2277 128X, Volume 3, Issue 10, October 2013.
[19] AlamJahaan , Dr T.N. Ravi, “A Capsulization OF Image Compression Aspects”, International Journal Of Advanced Research In Computer Science And Applications Issn 2321-872x Online Issn 2321-8932, Volume 3, Issue 7, July 2015.
[20] S.R. KodituwakkuEt. Al. “Comparison Of Lossless Data Compression Algorithms For Text Data”, Indian Journal Of Computer Science And Engineering, Vol 1 No 4 416-425
[21] Nour S. Bakr, Amr A. Sharawi,“DNA LosslessCompression Algorithms: Review”, American Journal of Bioinformatics Research 2013,3(3):72-DOI:10.5923/j.bioinformatics.20130303.04