iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** XI **Month of publication:** November 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue XI, November 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Digital Data Storage in DNA

A. D. Venkhat Balaji
MNM Jain Engineering College

Abstract: Demand for data storage is growing exponentially, but the capacity of existing storage media is not keeping up. Using DNA to archive data is an attractive possibility because it is extremely dense, with a raw limit of 1 Exabyte/mm3 (109 GB/mm3), and long-lasting, with observed half-life of over 500 years. This paper presents architecture for a DNA-based archival storage system. It is structured as a key-value store, and leverages common biochemical techniques to provide random access. We also propose a new encoding scheme that offers controllable redundancy, trading off reliability for density. We demonstrate feasibility, random access, and robustness of the proposed encoding with wet lab experiments involving 151 kB of synthesized DNA and a 42 kB random-access subset, and simulation experiments of larger sets calibrated to the wet lab experiments. Finally, we highlight trends in biotechnology that indicate the impending practicality of DNA storage for much larger datasets

I. INTRODUCTION

The demand for data storage devices is increasing day by day as more and more data is generated every day. Total information in digital format in the year 2012 was about 2.7 zettabytes. Presently devices such as optical discs, portable hard drives, and flash drives are used to store data. But silicon and the other non biodegradable materials used in data storage pollute the environment. Also, they are available in limited quantities. Thus, they would be exhausted one day. The linear density of digital storage device is 10 kb per square mm. Hence, newer technology is needed for data storage and archival process. As the data increases, the current data storage technology would not be enough to store data in future as data is growing every day. Even potentially important information can get lost due lack of storage space. One of the most common causes of data loss is accidental deletion of files without backup. Every day many people lose important data because of deleting files accidently because they do not have proper backup systems. Poor handling of the optical disk can cause data loss in them. Data loss can occur due to damage of hard drive. Mechanical damage to the hard drive is common as it contains a lot of fragile parts moving at very high speed. Hard drives can get damaged due to accidental drop of computers. Hard drives can get damaged if any liquid enters it. Liquids can cause damage to electronic parts of drive making it difficult to recover data. The hard disk can get damaged due to fire. Solid State Drive has a limited number of write cycles. Thus after write cycle limit, it is not possible to write data on them. A printed book has better life expectancy than best of the data storage method.

A. Importance of DNA as storage medium

1) High capacity
2) High data storage density
3) Withstand extreme environmental conditions
4) High memory space
5) Secure as invisible to human eye

B. Structure of DNA

1) DNA consisting of Adenine, Guanine, Cytosine and Thymine. (A,G, C and T)
2) Paired into nucleotide base pairs A-T and G-C
3) Single nucleotide can represent 2 bits of information.
4) High memory space due to 3D structure

C. Encoding Models

1) Microvenus project
2) Genesis project
3) PCR based encoding

Volume 4 Issue XI, November 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

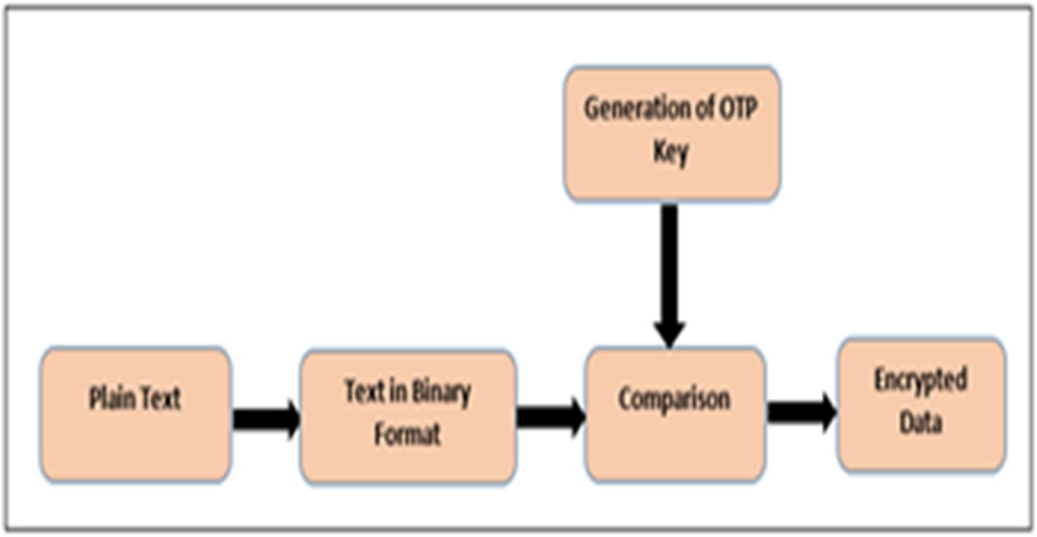
4) Alignment based encoding
5) Church and Goldman model
6) Rewritable RAM
7) Next generation sequencing model

D. Codes for encoding

1) The Huffman Code
2) The Comma Code
3) The Alternate Code
4) Comma Free Code
5) Improved Huffman Coding
6) Perfect Genetic Code

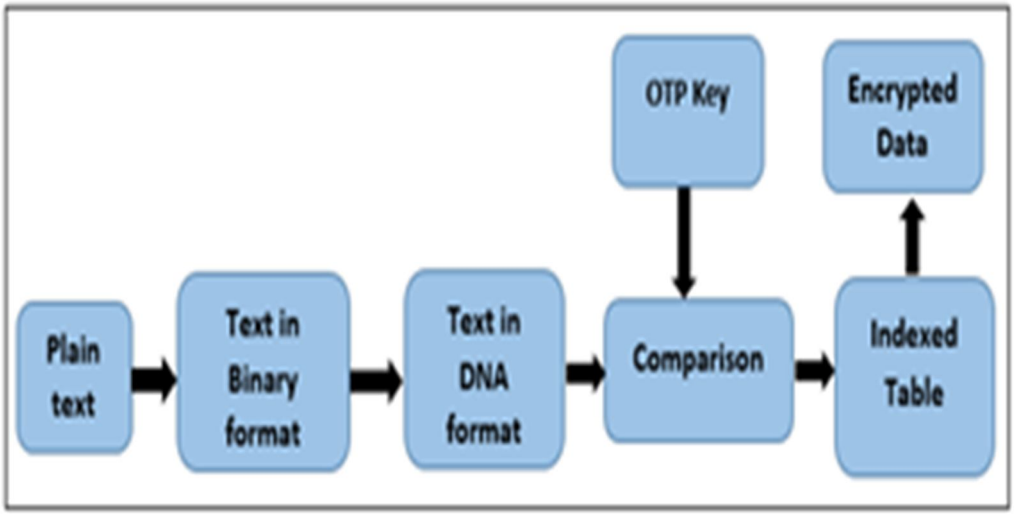
E. DNA secret writing algorithms

Steganography technique using DNA hybridization



F. DNA secret writing algorithms

Chromosomes DNA indexing



Volume 4 Issue XI, November 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

G. Comparison of codes used for encoding

	Huffman code	Comma code	Alternating code	Comma free code	Improved Huffman code
Features	Variable length code Used for short term and routine data storage	Fixed length codons separated by commas G:C to A:T pair ratio is 1:1 Creates fully synthetic DNA	Fixed length codons separated by commas G:C to A:T pair ratio is 1:1 Creates fully synthetic DNA	Establishes an automatic codon frame	Easy pattern recognition using specially designed primers
Advantages	Economical Unambiguous	Message DNA has isothermal melting temperature Detects errors Unambiguous Comprise of clear reading frame Protection against insertion and deletion mutations Suitable for long term storage	Message DNA has isothermal melting temperature Detects errors Suitable for long term storage	Economical Error detecting to certain extent	Economical Unambiguous Stores texts, images and music in DNA Error detection of mutations due to frame shifts Possess error correction mechanism as it can recover data from damaged DNA
Disadvantages	Not applicable to symbols and numbers Not appropriate for long term storage No reading frame	Not economical	No automatic reading frame Not economical	Does not construct fully artificial DNA	Expensive

II. DATA PREPARATION

A. Context Information Generation

Currently there are many compression methods that require good context in order to achieve a good compression ratio. One of them is Burrow Wheeler transform. BWT can achieve good compression ratio provided that there is a sufficient context which is formed by frequent occurrence of symbols with same or similar prefixes. Maximizing the context leads to better compression ratio. The Burrow Wheeler algorithm is based on the frequent occurrence of symbol pairs in similar context and it uses this feature for obtaining long strings of the same characters. These strings can be transformed to another form with move to front (MTF)

transformation.

B. Compression of Text File

We used statistical compression method to compress the data obtained after transformation. The chosen statistical compression scheme was Huffman encoding. Input consists of alphabet A and set W represented in equation (1) and (2) respectively. Output is a set of binary sequence in equation (3), which must satisfy the goal (4) for all the codes with the given condition.

$$A = \{a1, a2\ldots, an\} \quad (1)$$

$$W = \{w1, w2\ldots, wn\} \quad (2)$$

$$wi = \text{weight } (ai\ ),\ 1<i< n$$

$$C\ (A, W) = \{c1, c2\ldots, cn\} \quad (3)$$

III. MAPPING FUNCTION

A. Mapping Table

Mapping table consists of binary bits and nucleotides. Binary value is represented as 0 and 1. Nucleotides are represented as A, C, G and T. Four binary bits are represented by two nucleotide base pairs resulting in sixteen such combinations as shown in Mapping Table. The reason for choosing four bits for two nucleotides is that the output of Huffman encoding here is Hexadecimal value (radix =16). So we need sixteen such combinations to represent this in binary and then nucleotides.

Binary - nts	Binary - nts	Binary - nts	Binary - nts
0000 - AA	0100 -AC	1000 -AG	1100 -AT
0001 -CA	0101-CC	1001-CG	1101-CT
0010-GA	0110-GC	1010-GG	1110-GT
0011-TA	0111-TC	1011-TG	1111-TT

B. Encryption

The encoded message must be encrypted in order to maintain its security. For this purpose One time pad encryption is used. The requirement for one time pad is that the number of bits of random key must be the same length as of the message to be encoded. The encryption is processed character by character. The secrecy property of the encrypted message depends upon the generated random pad and the decryption of the message is impossible without knowing the true random key and makes it mathematically unbreakable.

IV. MESSAGE ENCODING AND RETRIEVAL

We have implemented the data encoding in nucleotides by integrating the trans-formation algorithm with statistical compression scheme. Here we have demonstrated our encoding and message retrieval scheme on small text: OPERATION BARBAROSSA. The first step was to perform Burrow wheeler transform and move to front transform on the original text. This was done to generate better context information and obtain high compression ratio. The security of the encoded message was maintained by encryption method. The encryption method used was One Time Pad where a randomly generated binary strand was XORed with the binary strand obtained from Huffman en-coding. We used a random function generator for generating the random binary sequence. Only in the last step, Huffman encoding method was introduced which compressed the original text message to a much smaller size. The next step toward message encoding was to use mapping table. The generated binary strand obtained after Huffman encoding was mapped to nucleotides according to the mapping table. Second phase of our work was to convert the encrypted binary strand into nucleotide sequence. Although many other mapping functions can be used, but for our convenience we used two nucleotides to represent four binary bits, as hexadecimal (radix =16) value is being converted to four bit binary representation and thus leading to formulation of original text message in form of nucleotide sequence. The decoding of message can be performed by reversing the encoding scheme. This is explained in Figure 1. This nucleotide sequence can be artificially synthesized and inserted into the host to maintain the attributes of hereditary media and durable data storage for intensive period of time. We have not proceeded in

Volume 4 Issue XI, November 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

implementing the biological protocols to insert the sequence in genome of bacteria.

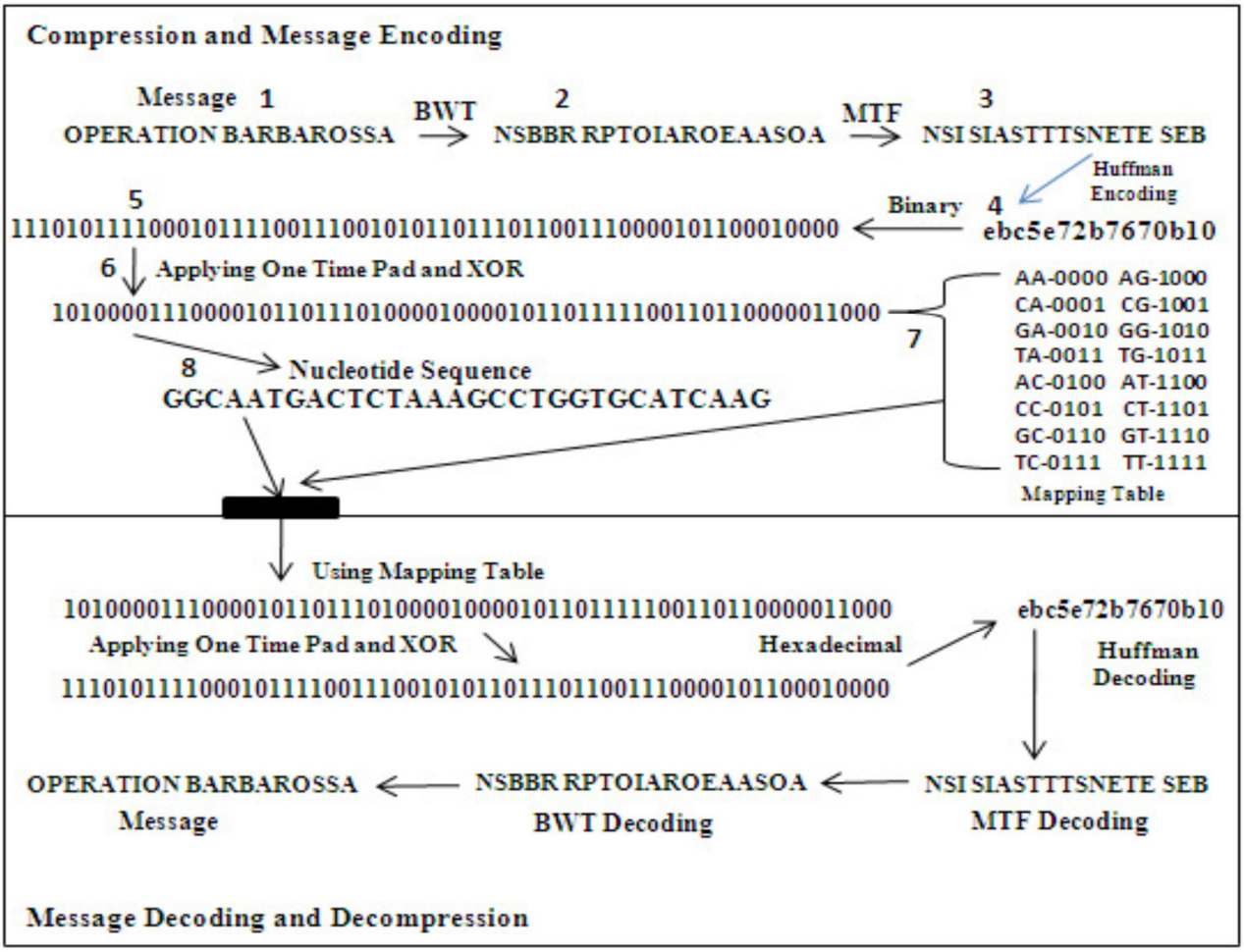
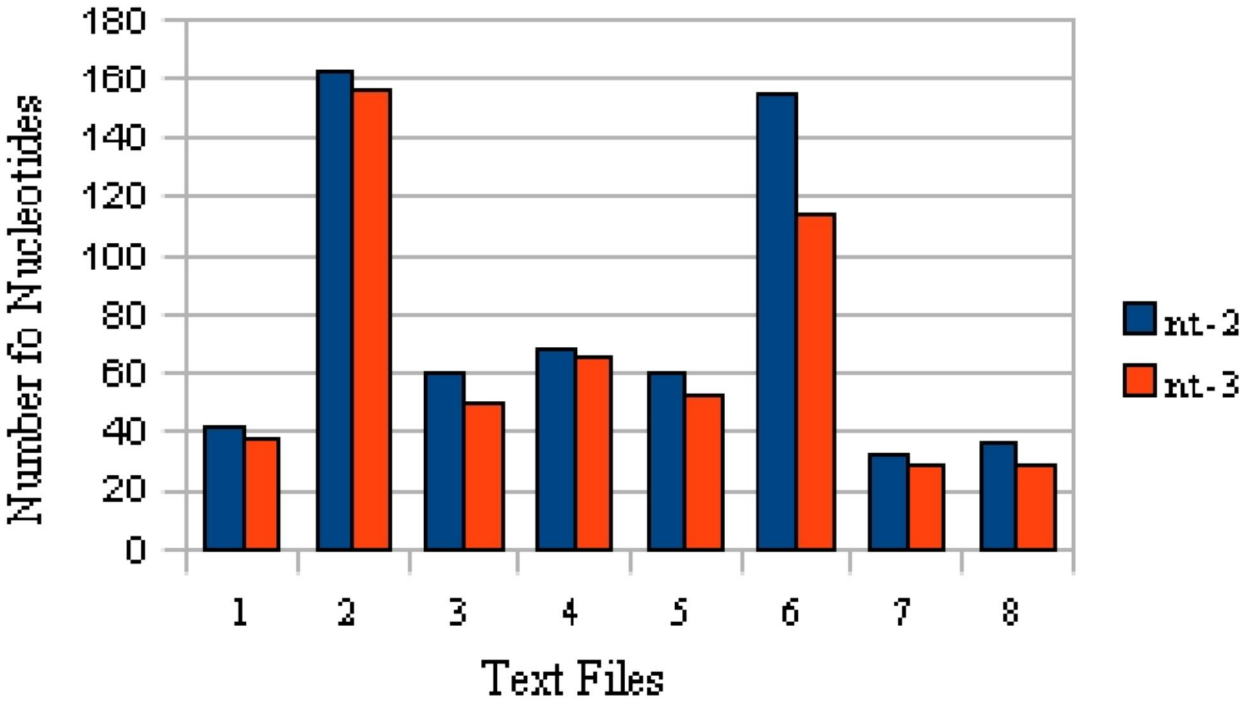


Fig. 1. Message encoding and decoding scheme in nucleotides.

V. PERFORMANCE

The suggested method was compared with two different methods on a set of les. The les that we used for testing is available at http://testdata.idlecool.net. This is a collection of single line texts. The original form was used for compression and encoding scheme. The first method was comparison of nt-1 and nt-3, where nt-1 represents the number of nucleotides used to represent the original text message after converting the text message to binary and mapping it with nucleotides, and nt-3 represents the number of nucleotides obtained after encoding the nucleotide sequence after performing transformation and then applying compression algorithm on the same text message. In this algorithm, Burrow-Wheeler Transformation and Move to Front transform was applied to the text message. The comparison in the second method was used for demonstrating the importance of context information generation using transformation algorithm. Here we compared nt-2 and nt-3, where nt-2 represents the number of nucleotides obtained after encoding then nucleotides without applying transformation algorithm and nt-3 with transformation algorithm on the same text le. The compression efficiency was tested with many tests over several les. The size of text les chosen varied from 140 Bits to 700 Bits. Experimental result showed that the maximal compression efficiency was achieved by applying transformation in the first step. This transformation generates better context information needed to compress text les of very small size (1000 Bytes). Result for encoding nucleotides without performing transform is depicted in the Figure 3. This shows that transformation prior to compression reduces the number of nucleotides to represent the same text message. The mean compression factor for the eight tested le was 2.076. As can be seen form results above, our algorithm for small message is better when incorporated with transformation algorithm. Even though, the difference in compression factor was 0.294. This is a step toward reducing the number of nucleotides for message strand and consecutively reducing the cost factor in artificially synthesizing the nucleotide strand for

encoding message.



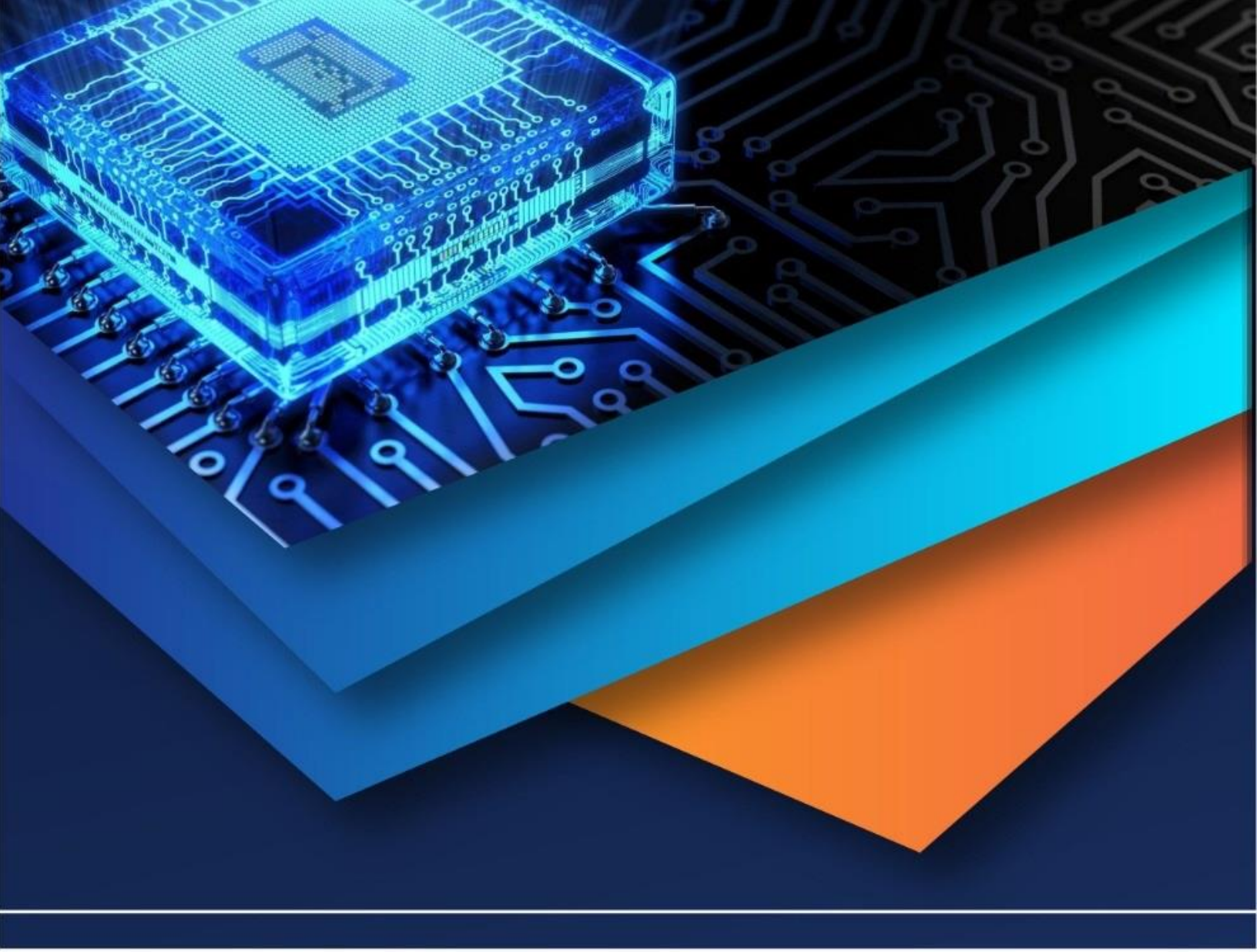
VI. CONCLUSION

This paper describes a data encoding method to achieve high volume data density by reducing the number of nucleotides. The primary focus of this study was to encode data for les of very small size. Data encoding method was performed into two steps. The first step was to compress the original text message. This was achieved using transformation and compression algorithm. Second step was introduction of mapping table, which finally maps the binary strand to nucleotide sequence. The contributions of this study are summarized as follows. First, while majority of previous experiments just mapped the binary strand to nucleotide sequence, this study reduced the number of nucleotides to represent the same in-formation, finally reducing the cost factor for artificially synthesizing nucleotide strand in laboratory. Second, this study uses transformation on original text message prior to compression to achieve better context information, resulting in a compression factor of 2.37. This scheme may be useful for many applications which need to store small message for long period of time, e.g. military implications , signatures of living modified organism (LMOs) and as valuable heritable media. Future work may focus on modification of transformation algorithm and designing other mapping function for encoding nucleotide sequence.

REFERENCES

[1] "Extracting Value from Chaos" (IDC, Framingham, MA, 2011), http://www.emc.com/collateral/analyst-reports/idc-extracting-value-fromchaos-ar.pdf.
[2] Information on materials and methods is available on Science Online.
[3] C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland, Long-term storage of information in DNA. Science 293, 1763 (2001). doi:10.1126/science.293.5536.1763c Medline
[4] E. M. LeProust et al., Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Res. 38, 2522 (2010). doi:10.1093/nar/gkq163 Medline
[5] SanDisk, SanDisk Develops Smallest 128Gb NAND Flash Memory Chip (SanDisk, Milipitas, CA 2012, http://www.sandisk.com/about-sandisk/pressroom/press-releases/2012/sandisk-develops-worlds-smallest-128gb-nandflash-memory-chip).

Volume 4 Issue XI, November 2016
ISSN: 2321-9653



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY