Big Data into DNA storage

[1]Anushree Raj, [2]Rio G.L. D'Souza

[1]Assistant Professor, [2]Professor
[1]M.Sc. Big Data Analytics Department, [2] Computer Science and Engineering Department,
[1,]St Agnes College Autonomous Mangalore, Karnataka, India, [2] St Joseph Engineering College Mangalore, Karnataka, India

***Abstract*: DNA is the world's oldest data storage device. The technology to read and write DNA has become common place since bacteria were first genetically engineered in 1973. While it's possible to store petabytes of data in a microscopic space, its always worthwhile to store information as DNA, rather than on hard drives or magnetic tape. "DNA is remarkable: just one gram of DNA can store about a petabyte's worth of data, and that's with the redundancy required to ensure that it's fully error tolerant. It's estimated that we can put the whole internet into the size of a van. The BIG Data Centre at Beijing Institute of Genomics (BIG) of the Chinese Academy of Sciences provides freely open access to a suite of database resources in support of worldwide research activities in both academia and industry. This topic "Big Data into DNA storage" demonstrates the various steps involved in storing huge digital data into a DNA. The experimental work is done to validate the proposed algorithms.**

***IndexTerms*: Big data storage, DNA storage, DNA Encoder, DNA Decoder.**

I. INTRODUCTION

The term Big Data is usually used to describe huge amount of data that is generated by humans from digital media. Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale [1]. The term DNA sequencing refers to sequencing methods for determining the order of the nucleotide bases - adenine, guanine, cytosine, and thymine - in a molecule of DNA. However, there are several technological constraints for writing (synthesizing), storing, and reading (sequencing) DNA [2]. DNA digital data storage is defined as the process of encoding and decoding binary data to and from synthesized DNA strands [3]. DNA as a storage medium is extremely effective. It is compact, biodegradable, and consumes very little energy. Today it is used to propagate species, encode protein synthesis, and solve complex computational problems. DNA sequence has become the first molecular data for which the cost of storage has become a significant proportion of the overall cost of generation and analysis [4]. This paper proposes four algorithms to encode and decode files into DNA sequence, which is involves synthesis, storage and sequencing of DNA library. The proposed algorithms are tested for different types of data and results are compared with different types of data file with respect to its sizes.

II. BIGDATA STORAGE IN DNA

Era of DNA computing begins with the identification of limitations in electronic computers. The volume of data that can be stored in an electronic computer and the speed thresholds that can be reached which is governed by the physical characteristics of computers are the main limitations identified in big data storage[5]. DNA computer addresses the above-mentioned limitations through solving computational problems engaging molecule manipulations while discovering natural computational models leading to the bigdata storage and bigdata analytics in DNA. The concept of storing data in DNA molecules is well discussed [6]. Advantages offered by DNA computing include consuming significantly less energy than the electronic computers. Energy consumed by DNA computers is billion times comparatively less than other electronic computers. The storage space needed to store information is less than trillion times over electronic computers. Furthermore DNA computers offer parallelism at a high level. Millions and trillions of molecules per form chemical reactions parallel [7].

III. ALGORITHMS AND TECHNIQUES USED

The proposed algorithms are implemented using java code. The algorithms are run on windows and every encoded and decoded file is stored in txt format. The process involves set of algorithms and techniques to encode and decode different types of data files. Base64 encoding algorithm is used to convert any file into its base64 format. DNA encoding algorithm converts the base64 file into a DNA sequenced file. Techniques like DNA synthesis [8, 9], DNA library storage [10, 11] and DNA sequencer [12, 13] are used to save and retrieve the chemical DNA sequence to and from its storage library. DNA sequences that are not intended to undergo processing by cellular machinery, but are just chemical messages such as DNA computing solutions, will be called "chemical DNA" [14]. DNA decoding algorithm is used to decode the DNA sequence into base64 format. Base64 decoding algorithm is used to decode the base64 back to original string. The figure 1 shows the flow of encoding and decoding process.

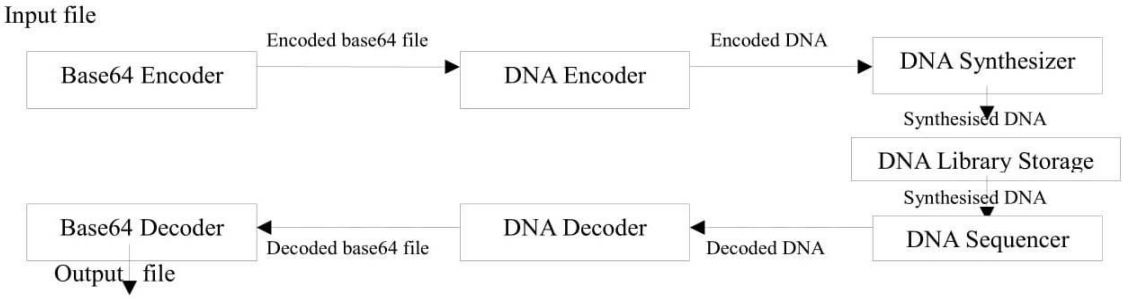

Figure 1. Steps involved in Encoding and Decoding process

3.1. Base64 Encoding Algorithm

Base64 encoding schemes are commonly used when there is a need to encode binary data that needs be stored and transferred over media that are designed to deal with textual data. This is to ensure that the data remains intact without modification during transport. Base64 is used commonly in a number of applications including email via MIME, and storing complex data in XML [15].

Algorithm Steps involved:

Step 1: Convert the actual data file into its ASCII format as bytes
Step 2: Convert each byte into its base 2.
Step 3: Group every 8 bits into buffer to produce a binary strings of 24 bit each.
Step 4: Split the binary strings into 6 bits each pad zeros to make up the final 6 bit codon.
Step 5: Convert every 6 bit into its base64 equivalent value.
Step 6: Store the result into a file
Input: Any file
Output: base64 encoded file

3.2. DNA Encoding Algorithm

The base64 file is further encoded to its binary equivalent and every pair of binary bits are converted to its DNA base for 00 represented as A, 01 represented as C, 10 represented as G and 11 represented as T.

Algorithm Steps involved:

Step 1: Convert the content of base64 file into binary format
Step 2: Split into groups of 8 bit each
Step 3: Refer the DNA base table to convert every pair of binary to its equivalent base. (A-00, C-01, G-10, T-11)
Step 4: Save the generated DNA sequenced strand into an file
Input: base64 format file
Output: DNA sequenced file

3.3. DNA Synthesis

There are four basic components required to initiate and propagate DNA synthesis. Polymerase Chain Reaction is a method for exponentially amplifying the concentration of selected sequences of DNA within a pool. A PCR reaction requires: substrates, template, primer and enzymes There are three basic codes for storing data in DNA which are Huffman code, comma code and the alternating code.[16].

3.4. DNA Library Storage

A DNA storage system consists of a DNA synthesizer that encodes the data to be stored in DNA, a storage container with compartments that store pools of DNA that map to a volume as shown in the figure 2 [13] .

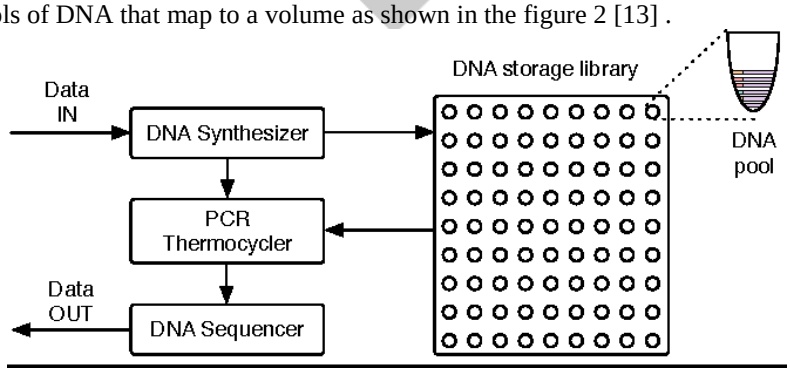

Figure 2

3.5. DNA Sequencing

DNA sequencing is the process of determining the sequence of nucleotides (As, Ts, Cs, and Gs) in a piece of DNA. Sequencing an entire genome (all of an organism's DNA) remains a complex task. It requires breaking the DNA of the genome into many smaller pieces, sequencing the pieces, and assembling the sequences into a single long "consensus" [17].

3.6. DNA Decoding Algorithm

The DNA sequence is now decoded to get back its original form. The DNA sequence is first converted to its binary form and split into 6 bits 4 codons. Each codon is further converted to its base64 format.
Algorithm Steps involved:
Step 1: Convert the DNA sequence into its binary equivalent referring to the base table (A-00, C-01, G-10, T-11)
Step 2: Split into groups of 6 bit each
Step 3: Convert the content of each 6 bit into its base 64 equivalent
Step 4: Save the generated base64 content into a file
Input: DNA sequenced file
Output: base64 format file

3.7. Base 64 Decoding Algorithm

The base64 file is now decoded back to its original form. Base 64 file is first converted into its binary form and then into its ASCII form of the original file.
Algorithm Steps involved:
Step 1: Convert each base64 character to its binary equivalent.
Step 2: Group 4 codons of 6 bit into a binary string of 24 bits.
Step 2: Split the 24 bit binary string into 3 bytes.
Step 3: Convert each byte into its ASCII equivalent
Step 4: Restore the result into its original file format
Input: base64 format file
Output: Decoded Original file

IV. EXPERIMENTAL WORK FOR PROPOSED ALGORITHMS

Steps	Process	Input
1	Input content	DNA DIGITAL DATA STORAGE IS DEFINED AS THE PROCESS OF ENCODING AND DECODING BINARY DATA TO AND FROM SYNTHESIZED DNA STRANDS
2	Base64 Encoding Algorithm	RE5BIGRpZ2l0YWwgZGF0YSBzdG9yYWdlIGlzIGRlZmluZWQgYXMgdGhlIHByb2Nlc3Mgb2YgZW5jb2RpbmcgYW5kIGRlY29kaW5nIGJpbmFyeSBkYXRhIHRvIGFuZCBmcm9tIHN5bnRoZXNpemVkIEROQSBzdHJhbmRz
3	DNA Encoding Algorithm (binary form)	01000100 01001110 01000001 00100000 01100100 01101001 01100111 01101001 01110100 01100001 01101100 00100000 01100100 01100001 01110100 01100001 00100000 01110011 01110100 01101111 01110010 01100001 01100111 01100101 00100000 01101001 01110011 00100000 01100100 01100101 01100110 01101001 01101110 01100101 01100100 00100000 01100001 01110011 00100000 01110100 01101000 01100101 00100000 01110000 01110010 01101111 01100011 01100101 01110011 01110011 00100000 01101111 01100110 00100000 01100101 01101110 01100011 01101111 01100100 01101001 01101110 01100111 00100000 01100001 01101110 01100100 00100000 01100100 01100101 01100011 01101111 01100100 01101001 01101110 01100111 00100000 01100010 01101001 01101110 01100001 01110010 01111001 00100000 01100100 01100001 01110100 01100001 00100000 01110100 01101111 00100000 01100001 01101110 01100100 00100000 01100110 01110010 01101111 01101101 00100000 01110011 01111001 01101110 01110100 01101000 01100101 01110011 01101001 01111010 01100101 01100100 00100000 01000100 01001110 01000001 00100000 01110011 01110100 01110010 01100001 01101110 01100100 01110011
4	DNA Encoding Algorithm (DNA Sequence)	CACA CATG CAAC AGAA CGCA CGGC CGCT CGGC CTCA CGAC CGTA ACAA CGCA CGAC CTCA CGAC AGAA CTAT CTCA CGTT CTAG CGAC CGCT CGCC ACAA CGGC CTAT ACAA CGCA CGCC CGCG CGGC CGTC CGC CGCA AGAA CGAC CTAT AGAA CTCA CGGA CGCC ACAA CTAA CTAG CGTT CGAT CGCC CTAT CTAT ACAA CGTT CGCG CAAA CGCC CGTC CGAT CGTT CGCA CGGC CGTC CGCT AGAA CGAC CGTG AGCA AGAA CGCA C10CC CGAT CGTT CGCA CGGC CGTG CGCT AGAA CGAG CGGC CGTG CGAC CTAG CTGC AGAA CGCA CGAC CTCA CGAC AGAA CTCA CGTT AGAA CGAC CGTG CGCA AGAA CGCG CTAG CGTT CGTC AGAA CTAT CTGC CGTG CTCA CGGA CGCC CTAT CGGC CTGG CGCC CGCA AGAA CACA CATG CAAC AGAA CTAT CTCA CTAG CGAC CGTG CGCA CTAT

5	DNA Synthesis, DNA Library Storage	
6	DNA Sequencing	CACA CATG CAAC AGAA CGCA CGGC CGCT CGGC CTCA CGAC CGTA ACAA CGCA CGAC CTCA CGAC AGAA CTAT CTCA CGTT CTAG CGAC CGCT CGCC ACAA CGGC CTAT ACAA CGCA CGCC CGCG CGGC CGTC CGC CGCA AGAA CGAC CTAT AGAA CTCA CGGA CGCC ACAA CTAA CTAG CGTT CGAT CGCC CTAT CTAT ACAA CGTT CGCG CAAA CGCC CGTC CGAT CGTT CGCA CGGC CGTC CGCT AGAA CGAC CGTG AGCA AGAA CGCA C10CC CGAT CGTT CGCA CGGC CGTG CGCT AGAA CGAG CGGC CGTG CGAC CTAG CTGC AGAA CGCA CGAC CTCA CGAC AGAA CTCA CGTT AGAA CGAC CGTG CGCA AGAA CGCG CTAG CGTT CGTC AGAA CTAT CTGC CGTG CTCA CGGA CGCC CTAT CGGC CTGG CGCC CGCA AGAA CACA CATG CAAC AGAA CTAT CTCA CTAG CGAC CGTG CGCA CTAT
7	DNA Decoding Algorithm (Binary Sequence)	01000100 01001110 01000001 00100000 01100100 01101001 01100111 01101001 01110100 01100001 01101100 00100000 01100100 01100001 01110100 01100001 00100000 01110011 01110100 01101111 01110010 01100001 01100111 01100101 00100000 01101001 01110011 00100000 01100100 01100101 01100110 01101001 01101110 01100101 01100100 00100000 01100001 01110011 00100000 01110100 01101000 01100101 00100000 01110000 01110010 01101111 01100011 01100101 01110011 01110011 00100000 01101111 01100110 00100000 01100101 01101110 01100011 01101111 01100100 01101001 01101110 01100111 00100000 01100001 01101110 01100100 00100000 01100100 01100101 01100011 01101111 01100100 01101001 01101110 01100111 00100000 01100010 01101001 01101110 01100001 01110010 01111001 00100000 01100100 01100001 01110100 01100001 00100000 01110100 01101111 00100000 01100001 01101110 01100100 00100000 01100110 01110010 01101111 01101101 00100000 01110011 01111001 01101110 01110100 01101000 01100101 01110011 01101001 01111010 01100101 01100100 00100000 01000100 01001110 01000001 00100000 01110011 01110100 01110010 01100001 01101110 01100100 01110011
8	Base64 Decoding Algorithm	RE5BIGRpZ2l0YWwgZGF0YSBzdG9yYWdlIGlzIGRlZmluZWQgYXMgdGhlIHByb2Nlc3Mgb2YgZW5jb2RpbmcgYW5kIGRlY29kaW5nIGJpbmFyeSBkYXRhIHRvIGFuZCBmcm9tIHN5bnRoZXNpemVkIEROQSBzdHJhbmRz
9	Output Decoded file	DNA digital data storage is defined as the process of encoding and decoding binary data to and from synthesized DNA strands

V. EXPERIMENTAL RESULTS

The code is tested for different types of files of various sizes. The encoding and decoding is possible for very huge sized data and for most of the file formats. The resultant decoded file seems to be error free assuming the DNA synthesis and sequencing is done accurately. The encoded files are temporarily stored in the server and deleted once the DNA sequence is stored in the library. Decoder is used to retrieve the data back into its original format. Table 1 shows the sizes of different types of files in the process of encoding and decoding.

S. No.	File Type	Original File size	Base64 file size	DNA file size	Decoded file size
1	Doc file	8.6MB	9.3MB	7.4MB	8.6MB
2	Pdf file	144.5MB	153.3MB	132.8MB	144.5MB
3	Jpg file	211.6MB	226.8MB	201.4MB	211.6MB
4	Mp3 file	643.4MB	737.4MB	593.4MB	643.4MB
5	Mp4 file	11346.8MB	12472.5MB	10034.9MB	11346,8MB

Table 1

VI. CONCLUSSION

The current rate of data explosion is increasing at a very high rate, DNA storage becomes an absolutely key for data storage because of its low maintenance cost, high data density, eco friendliness and durability. DNA storage techniques show massive progress. The number of publications related to various DNA based models and techniques increases tenfold annually [18]. Presented methods literally convert each and every smallest possible information into DNA form. Thus this method is highly applicable for handling and storage of massive amounts of various types of data. The proposed algorithms works good for various Big data file formats. Since java code is used, the compilation for huge files becomes an overhead with respect to time. The proposed mechanism can further be modified by using latest coding languages which are more suitable for huge data file which works on low compilation time.

References

[1] Sagiroglu, S.Sinanc, D.,‖Big Data: A Review‖,2013, 20-24.

[2] K. Reinert and D.H. Huson, Bioinformatics Support for Genome-Sequencing Projects. In T. Lengauer, editor, Bioinformatics- From Genomes to Therapies, Vol. 1., pages 25-56, Wiley-VCH Verlag, 2007.

[3] Bentley DR. 2006. Whole-genome resequencing. Curr.Opin.Genet.Dev. 16:545–52

[4] Efficient storage of high throughput sequencing data using reference-based compression. Markus Hsi-Yang Fritz, Rasko Leinonen, Guy Cochrane, et al. Genome Res. published online January 18, 2011.

[5] Bakshi, K., 2012. Considerations for big data: Architecture and approach. In: 2012 IEEE Aerospace Conference, Big Sky Montana. pp. 1-7.

[6] NEXT-GENERATION DIGITAL INFORMATION STORAGE IN DNA, GEORGE M. CHURCH, YUAN GAO, SRIRAM KOSURI, *SCIENCE* 28 SEP 2012: VOL. 337, ISSUE 6102, PP. 1628 DOI: 10.1126/SCIENCE.1226355

[7] Liu, H., D. Lin, and A. Kadir, A novel data hiding method based on deoxyribonucleic acid coding. Computers & Electrical Engineering, 2013. 39(4): p. 1164-1173.

[8] M. D. Matteucci and M. H. Caruthers. Synthesis of deoxyoligonucleotides on a polymer support. Journal of the American Chemical Society, 103(11):3185–3191, 1981

[9] S. Kosuri and G. M. Church. Large-scale de novo DNA synthesis: technologies and applications. Nature Methods, 11:499–507, 2014

[10] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney. Towards practical, high-capacity, lowmaintenance information storage in synthesized DNA. Nature, 494:77–80, 201

[11] G. M. Church, Y. Gao, and S. Kosuri. Next-generation digital information storage in DNA. Science, 337(6102):1628, 2012

[12] T. P. Niedringhaus, D. Milanova, M. B. Kerby, M. P. Snyder, and A. E. Barron. Landscape of next-generation sequencing technologies. Anal. Chem., 83:4327–4341, 2011.

[13] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss. A DNA-based Archival Storage System. In ASPLOS, 2016.

[14] Marvin H. Caruthers;The Chemical Synthesis of DNA/RNA: Our Gift to Science. January 11, 2013 The Journal of Biological Chemistry, 288, 1420-1427

[15] Base64 Character Encoding and Decoding Modeling Isnar Sumartono1 , Andysah Putera Utama Siahaan2 , Arpan3 Faculty of Computer Science, Universitas Pembangunan Panca Budi Jl. Jend. Gatot Subroto Km. 4,5 Sei Sikambing, 2012, Medan, Sumatera Utara, Indonesia

[16] POLYMERASE CHAIN REACTION: AN EMERGING TOOL FOR RESEARCH IN PHARMACOLOGY RISHI SHARMA, MANJEET SINGH, AJAY SHARMA. Indian Journal of Pharmacology 2002; 34: 229-236.

[17] A review of DNA sequencing techniques Lilian T. C. Franc:a1 , Emanuel Carrilho2 and Tarso B. L. Kist. Quarterly Reviews of Biophysics 35, 2 (2002), pp. 169–200. "2002 Cambridge University Press DOI: 10.1017/S0033583502003797 Printed in the United Kingdom.

[18] Review of Big Data Storage based on DNA Computing Hanadi Ahmed Hakami, Zenon Chaczko and Anup Kale. Published in 2015 Asia-Pacific Conference on Computer Aided System Engineering.