International Journal of Innovative Research in Computerand Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

|| Volume 8, Issue 4, April 2020 ||

A Review on DNA Data Storage

Punitha N[1], Savita Sheelavant[2]

Student, Department of MCA, Rashtreeya Vidyalaya College of Engineering, Bengaluru, India[1]

Assistant Professor, Department of MCA, Rashtreeya Vidyalaya College of Engineering, Bengaluru, India[2]

ABSTRACT: In the present day, the data is increasing day by day from the different sources such as social media, health management, etc... Presently devices such as optical HDD and etc are used to store data. All these non-biodegradable material used in data storage pollute the environment. As the information builds, the present information stockpiling innovation would not be sufficient to store information in future as information is developing each day. To backup the data, various hard drives and big data centers will be used to harvest the important data.

Considering the way fossil bones save hereditary material for a very long time, specialists gave their consideration towards utilizing deoxyribonucleic acid (DNA) as a storage medium. Hence DNA can be used to store the data for the longest period in any domain. This paper reviews about the DNA as the storage medium and the process for the storing the data into the DNA.

KEYWORDS*:* DNA Data Storage, Improved Huffman Code, Steganography Technique

I.INTRODUCTION

DNA or Deoxyribonucleic Acid is defined long molecule where our unique genetic code will be stored. DNA has an unbelievable storage capacity. The aftereffect of ongoing inquires about states that all the data in the whole web could be situated in a gadget which is lesser than unit cubic inch. DNA is witnessed as the optimal storage medium in this regard fundamentally because DNA consisting of adenine, guanine, cytosine and thymine (A,G,C and T) already paired into two nucleotide base pairs A-T and G-C which can be used for storing information in the form of binary code, instead of using 1's and 0's by the computer to store data.

DNA has been identified as the potential medium for data storage due to its vast storage capacity, high data density, sustaining to extreme environmental conditions, and so forth.DNA is used successfully in confidential writing because of the protection features DNA provides. DNA cryptosystems are more secure due to the huge size of the OTP key used for encryption. Breaking the algorithm would be incredibly difficult without understanding the organism's primary sequences and scientific requirements. Running time of cryptographic systems is less.

Since DNA can store the large amount for the longest period, it can be used to store the entire data of any industry or company or an institute. Figure 1 explains, when data is ready to get stored in DNA, it will be encoded using some algorithms like Huffman code, the comma Code. If any error occurs during the encryption then the error will corrected using insertions, deletion and substitutions.

Once the error is corrected DNA secret writing algorithm will be applied to write the data into DNA. Considering the way fossil bones save hereditary material for a very long time, specialists gave their consideration towards utilizing deoxyribonucleic acid (DNA) as a storage medium. Hence DNA can be used to store the data for the longest period in any domain

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

|| Volume 8, Issue 4, April 2020 ||

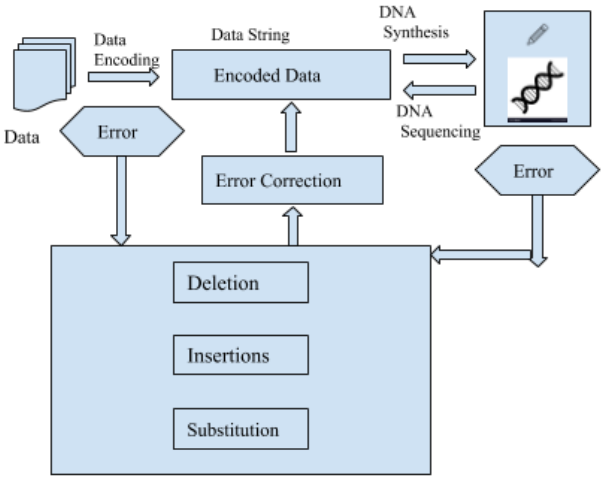

Figure 1: Block Diagram of the DNA Data Storage

II.RELATED WORK

Literature Survey

The two biochemical constraints, maximum homoploymer run limit and balanced GC content limit were addressed by Yixin Wang, Md. Noor-A-Rahim and Erry Gunawan. They proposed a run-length limited code which is novel content-balanced with an effective code construction method that generates short DNA sequences that satisfies both constraints at a time [1]. Douglas Carmean and Luis Ceze have explored the implications of DNA as storage medium trend in computer architecture. They proposed a more design which is ambitious hybrid–electronic, which uses a molecular form of near-data processing for massive parallelism. They even presented a model that demonstrates the feasibility of these systems in the near future [2]. Reza A. Ashrafi and Ali E. Pusane talked about how the data retention capabilities drastically decline with the increased need for long term and large storage space in the solid-state based memory devices. They reviewed two promising storage technologies i.e. Solid-state based memory devices and Biotechnological based DNA storage and their potential future trends are discussed [3].

Organick L and Ang S encoded and stored 35 distinct files (over 200 MB of data), in more than 13 million DNA oligonucleotides, and show that each file individually can be recovered and with no errors, using a random access approach. A large library of primers has been designed and validated that enable individual recovery of all files stored within the DNA. An algorithm has been built up that enormously decreases the sequencing read inclusion required for mistake free translating by expanding all arrangement read data. These developments demonstrate a functional, large-scale DNA data storage and retrieval system [4].

Reinhard Heckel and Ilan Shomorony studied a simple about which exhibits the fundamental limits and tradeoffs of DNA based storage systems, motivated by the current technological constraints on DNA synthesis and sequencing. Their model captures two important aspects of the DNA storage systems. First aspect is many short DNA molecules are written with the data in unordered way and second aspect is the data is read by randomly sampling from this DNA pool. In this model they characterized the storage capacity and show that simple index-based coding is optimal [5].

Meinolf Blawat and Klaus Gaedke have built up a productive and strong forward blunder amendment conspire which is adjusted to the DNA channel. The design of the needed DNA channel model was designed on data from a proof-of-concept conducted 2012 by a team from the Harvard Medical School. The proposed forward error correction scheme was able to cope with all types of error of the current processes of DNA synthesis, amplification and sequencing, e.g. insertion, deletion, and swap errors [6]. R. Pragaladan and S. Sathappan have talked the purpose of using cloud infrastructure with DNA structure using bio-molecular method has complex form that makes difficult to analyze and therefore improving the data confidentiality [7].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

|| Volume 8, Issue 4, April 2020 ||

Naveen Goela and Jean Bolot focused on the systems level design for error correction is presented for encoding movies and digital information in DNA storage. Their goal was to decode the source from the DNA reliably even in the presence of diverse errors introduced by DNA Synthesis, PCR amplifications and DNA sequencing processes [8]. S. M. Hossein Tabatabaei Yazdi and Han Mao Kiah provided an overview of approaches to DNA-based storage system design and of accompanying synthesis, sequencing and editing methods. They also introduced and analyzed a suite of new constrained coding schemes for both archival and random access DNA storage channels. Their contributions is in the work is the construction and design of sequences over discrete alphabets that avoid pre-specified address pattern, have balanced base content, and exhibit other relevant substring constraints. These schemes adapt the stored signals to the DNA medium to reduce the inherent error-rate of the systems [9].

Hanadi Ahmed Hakami, Zenon Chaczko and Anup Kale have talked about the need of noteworthy scaling down in the information approached may be saved in the most recent decade. In their paper, they have described and reviewed the every research on DNA storage, their advantages and disadvantages, their technologies and how they would become a practice in the future. They also proposed an approach is proposed a simple method to store the data into DNA. They experimented a work to validate the proposed approach result clearly show advantages merits of proposed method [10].

Algorithms

Here are some algorithms which are used in the process of the storing the data into the DNA.

- **The Huffman Code Algorithm**

The principle of varying the lengths of symbols used for representing a character is used by this code. The lowest number of symbols is assigned with the most recurrently appearing character in a text while the least recurrently appearing character is assigned the most number of symbols. Deploying this principle will be helpful in developing an economical code. Around 2.2 characters is the average code length in The Huffman Code. This is the least average codon length achieved.

Comprising of only one way in which the encrypted message can be read once the starting point of the message is mentioned will be helpful in achieving the unambiguity of the code.

Drawbacks of the Huffman code include not outfitting for numbers and symbols. This is mainly because of the frequency of showing these symbols is highly dependent on the text which receives the fact that they are unable to be included in formulating the Huffman code. Secondly, because of the fact that when different length codons assembled together it might not reveal a pattern, the Huffman code is not suitable for long term storage.

- **The Comma Code**

In the comma code approach a single base Guanine (G) is considered as the comma. Codons of 5-base length are separated from each other using G. % base codons consists of other three bases A,C, and T, further more limited to single A:T base pair and two G:C pairs. The C of second G: C pair is always positioned in the upper strand.

Comprising of isothermal softening temperature is the upside of the structure of the message DNA using this plan. Prevailing aspect of the comma code is the perusing edge of six codons including G, the comma, which isn't accomplished by different codes. This assists with distinguishing an unmistakable perusing outline without the need to make reference to a beginning stage. Protection mechanism from insertion and deletion mutation is also guaranteed by this approach which makes other codes much more complex.

Disadvantage of the comma code is, since is repeats the comma-base G to create an automatic reading frame, comma code is not economical.

- **Comma-Free Code**

It is also known as prefix free code. This includes fixed length base casings without commas to isolate the edges. And it uses a method for automatic frame detection.. Comma-free code does not consist of identical four base pairs which is the only way of hindering from natural DNA sequences. These codons can be interpreted easily in one way, and also support mechanisms for error detection.

While comma-free code is robust and error correction works to correct against small-scale losses like DNA point mutations, it does not have the ability to recover broken data when a large DNA segment is removed from the DNA region encoded by the data.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

|| Volume 8, Issue 4, April 2020 ||

- **Improved Huffman Coding Scheme**

Each approach utilized for data stockpiling in DNA varies in the prudent utilization of nucleotides. Here Ailenberg and Rotstein utilize the standards of Huffman coding to characterize DNA codes for the whole console, for obvious data coding. This conquers the downside of the Huffman code, being restricted distinctly to the letters of the letters in order. This depends on a development of a plasmid library with uniquely structured ground works implanted alongside the message for quick recovery. Record plasmid just contains insights concerning the structure of the data library. A decent encoding plan ought to have affordable utilization of nucleotide per character which is about 3.5 here.

Other coding plans have low base-to-character proportion however is constrained to bring down number of characters, for example, the English letter set. DNA was embedded into living life forms and they are liable to losing data because of breakage by change, addition, and erasure. Thus, this methodology is an answer for this issue as it can recoup information of harmed DNA. Therefore this method over comes the drawback of the comma-free code. This is additionally ready to recognize any casing shift because of transformation or mistakes in sequencing. This technique utilizes remarkable preliminary plan utilizing plasmid DNA libraries.

- **Steganography Technique using DNA Hybridization**

Advantages offered by the structure of DNA for vast storing capabilities and parallel molecular computation are effectively used by this method. An algorithm for hiding data in DNA in a digital form is identified and created by this approach. Keys generated by One Time Pad (OTP) are used as encryption key. This key is utilized simply just a single time for precisely one message. User will destroy the used pad after encryption. The process of Steganography technique is shown pictorially in the figure 2.

Because it is difficult for two complementary strands to combine together, DNA hybridization is a slower process at the beginning. But later this is a rapid process. This can be adequately used in looking and equal calculation. Time consumption and expansiveness are the restrictions for this process at present.

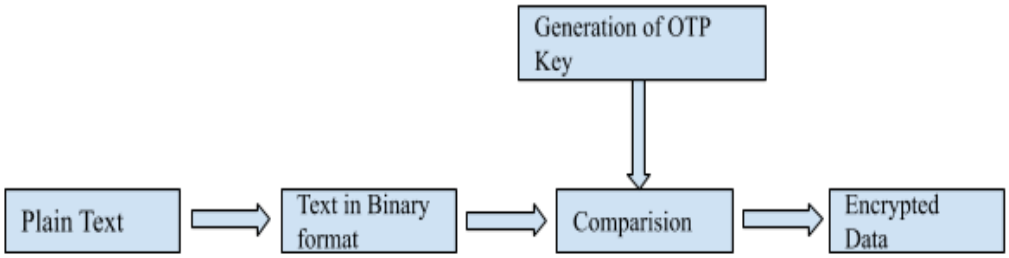

Figure 2: Block Diagram of the Steganography Technique

Analysis

Advantages of the DNA as the storage medium includes high capacity, high data storage density, high memory space, data will be secured as its invisible to human eyes, it can store the data for the longest period, and it uses the power very effectively. There are even some major issues regarding data storage in DNA such as, retrieval process is slower than that of personal computers, very high cost of productions, mutations in DNA and process of enhancing, rebuilding, deciphering, encoding and sequencing takes fundamentally a larger number of times than their regular partners.

The most of the encoding algorithms have some disadvantages. So the Improved Huffman code could be used efficiently instead of some other algorithms since it overcomes the drawbacks of the other algorithms like Huffman Code, Comma Code. The DNA was inserted into living organisms and they are subject to losing information due to breakage by mutation, insertion, and deletion. Hence, this approach is a solution to this problem as it is able to recover data of damaged DNA. Therefore this method overcomes the drawback of the comma-free code.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

|| Volume 8, Issue 4, April 2020 ||

III.CONCLUSION

DNA-based storage has the potential to be the ultimate storage solution in the educational institutes, industries and companies, which is extremely dense and durable. While this isn't viable yet because of the present condition of union and sequencing, the two innovations are improving at an exponential rate with progresses in the biotechnology business. Many algorithms like Huffman Code, Comma Code, etc., are used to encode the data. DNA Secret Writing algorithm Steganography Technique is used write the encoded data into the DNA, this process is known as DNA synthesis and sequencing. Using these technologies whole data of an institute can be stored in very small space for the longest period.

REFERENCES

1. Y. Wang, M. Noor-A-Rahim, E. Gunawan, Y. L. Guan and C. L. Poh, "Construction of Bio-Constrained Code for DNA Data Storage*"* in IEEE Communications Letters, vol. 23, no. 6, pp. 963-966, June 2019
2. D. Carmean, L. Ceze, G. Seelig, K. Stewart, K. Strauss and M. Willsey, "DNA Data Storage and Hybrid Molecular–Electronic Computing" in Proceedings of the IEEE, vol. 107, no. 1, pp. 63-72, Jan. 2019
3. R. A. Ashrafi, A. E. Pusane and S. S. Arslan, "Next-Generation data storage: Transistor and DNA" 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018
4. Organick, L., Ang, S., Chen, " Random access in large-scale DNA data storage", Nat Biotechnol **36,** 242–248 (2018).
5. R. Heckel, I. Shomorony, K. Ramchandran and D. N. C. Tse, "Fundamental limits of DNA storage systems" in IEEE International Symposium on Information Theory (ISIT), Aachen, 2017
6. Meinolf Blawat, Klaus Gaedke, Ingo Hütter, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin W. Pruitt, George M. Church, "Forward Error Correction for DNA Data Storage", Procedia Computer Science, Volume 80, 2016, Pages 1011-1022, ISSN 1877-0509
7. R. Pragaladan and S. Sathappan, "High Confidential Data storage using DNA structure for cloud environment" in International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, 2016
8. N. Goela and J. Bolot, "Encoding movies and data in DNA storage" in Information Theory and Applications Workshop (ITA), La Jolla, CA, 2016
9. S. M. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao and O. Milenkovic, "DNA-Based Storage: Trends and Methods" in IEEE Transactions on Molecular, Biological and Multi-Scale Communications, vol. 1, no. 3, pp. 230-248, Sept. 2015
10. H. A. Hakami, Z. Chaczko and A. Kale, "Review of Big Data Storage Based on DNA Computing" in Asia-Pacific Conference on Computer Aided System Engineering, Quito, 2015, pp. 113-117.