DNA technology for big data storage and error detection solutions: Hamming code vs Cyclic Redundancy Check (CRC)

Manar Sais, Najat Rafalia, Jaafar Abouchabaka

Department of Computer Science, Computer Research Laboratory LaRI, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

Abstract. There is an increasing need for high-capacity, high-density storage media that can retain data for a long time, due to the exponential development in the capacity of information generated. The durability and high information density of synthetic deoxyribonucleic acid (DNA) make it an attractive and promising medium for data storage. DNA data storage technology is expected to revolutionize data storage in the coming years, replacing various Big Data storage technologies. As a medium that addresses the need for high-latency, immutable information storage, DNA has several potential advantages. One of the key advantages of DNA storage is its extraordinary density. Theoretically, a gram of DNA can encode 455 exabytes, or 2 bits per nucleotide. Unlike other digital storage media, synthetic DNA enables large quantities of data to be stored in a biological medium. This reduces the need for traditional storage media such as hard disks, which consume energy and require materials such as plastic or metals, and also often leads to the generation of electronic waste when they become obsolete or damaged. Additionally, although DNA degrades over thousands of years under non-ideal conditions, it is generally readable. Furthermore, as DNA possesses natural reading and writing enzymes as part of its biological functions, it is expected to remain the standard for data retrieval in the foreseeable future. However, the high error rate poses a significant challenge for DNA-based information coding strategies. Currently, it is impossible to execute DNA strand synthesis, amplification, or sequencing errors-free. In order to utilize synthetic DNA as a storage medium for digital data, specialized systems and solutions for direct error detection and correction must be implemented. The goal of this paper is to introduce DNA storage technology, outline the benefits and added value of this approach, and present an experiment comparing the effectiveness of two error detection and correction codes (Hamming and CRC) used in the DNA data storage strategy.

Index Terms— *Big data, CRC, Error detection,* Energy, *Hamming, Synthetic DNA, Storage*

1 INTRODUCTION

Since the dawn of humanity, humans have continuously left traces of their writings, starting with rock art, the earliest known form of writing dating back 40,000 years. This art form includes sculptures and paintings on cave walls. Unfortunately, the remains of caves like Lascaux and Combarelles have not been fully preserved. The act of writing plays a paramount role in communication, enabling us to transmit and share our artifacts, knowledge, and the collective historical experiences of humanity and its surroundings with future generations. By 2025, the worldwide volume of produced data is projected to reach 250 zettabytes, and we find ourselves overwhelmed by an ever-growing mass of data. Current storage media is struggling to keep pace with this growth rate and lacks sufficient storage capacity, presenting significant challenges to current data centers and storage techniques. To address this issue, there has been a surge in research and development of Big Data platforms and systems, resulting in the emergence of various software products, tools, and database systems to meet the escalating demand for large-scale data storage and processing [1]–[4].

The term "data storage" refers to a collection of techniques and technologies used to gather and store digital information across various forms of media. Common examples include hard drives, flash drives, floppy disks, and SSDs. These storage options cater to individual users' needs, such as storing photos, documents, music, and more[4]. The volume of data has experienced a rapid surge due to the advent of big data, advanced analytics, and the proliferation of Internet of Things (IoT) devices [5]. These advancements have unlocked new opportunities in diverse industries, including scientific research, finance, commerce, and medicine. As a result, storage has become more crucial than ever in order to process the ever-increasing amounts of data. The main challenge presented by big data is finding a technology capable of efficiently storing and processing large volumes of data for analysis and information extraction. Several solution providers offer pre-packaged solutions to address big data challenges, such as Cloudera[6], HortonWorks[7], MapR[8], IBM Infosphere BigInsights[9], Pivotal HD[10], Microsoft HD Insight[11], and more[12]. The huge challenge that still exists is that all these current technologies are limited in terms of time tolerance or storage capacity.

Since the beginning of life on Earth, nature has found a solution to this issue on its own terms: it stores the information defining organisms in a unique sequence of four bases (A, T, C, G), which are located in a tiny molecule called deoxyribonucleic acid (DNA), this mode of information storage has continued for 3 billion years. Therefore, in order to meet the demand of storing more and more data, DNA as a storage medium has become an attractive choice, and it has numerous benefits over conventional storage media. [13].

According to Castillo[14], as far as storage is concerned, DNA has an amazing capacity for storing, all the data on the Internet may fit in a container that is less than one cubic inch in size. DNA storage can reach a theoretical density of 455 EB/g[15] and has a durable property of several centuries[16], [17]. From an environmental point of view, the manufacture of synthetic DNA can use environmentally-friendly synthesis methods, reducing the use of harmful chemicals and adopting sustainable practices. By developing more environmentally-friendly approaches to DNA synthesis, data storage in synthetic DNA can contribute to a more responsible use of resources.

Due to the characteristics of DNA data storage, errors that occur in the DNA chain pose a major challenge to the information encoding strategy and make the task of error detection and correction large and difficult. However, molecules can face errors during DNA storage, and these errors do not usually occur in traditional storage devices, such as deletion and insertion errors[18].

The structure of this essay is as follows: The research is introduced in Section 1. In Section 2, we discuss the theoretical background of this research by providing a general summary and overview of DNA storage technology. Section 3 aims to present synthetic DNA as an innovative form of big data storage, offering a promising solution for effectively preserving our data over the long term. The experimentation and results planned to encode DNA sequence data and decode DNA sequence data without error are shown in Section 4. We also compare the performance of the two-error detection and correction codes utilized in the experiment, CRC and Hamming codes.

2 RELATED WORK

Due to the increase in Internet traffic and the rapid development of the information industry, the amount of large and complex data is gradually increasing. Large volumes of data and various structures are being generated at an exponential rate, greatly beyond the capacity of conventional storage systems. Faced with the task of identifying, storing, and processing unstructured, noisy, dispersed, and heterogeneous data, the challenge of big data becomes more complex [19]. With the growing demand for superior technology that can efficiently store and process large amounts of data in a short period of time.

It is in this context that research began to turn, a few years ago, to DNA. DNA, deoxyribonucleic acid, exists in biological cells. It contains genetic information. Each DNA contains tens of thousands of genes, which determine the characteristics of each organism. With its enormous storage capacity, DNA can be considered a natural hard disk.

One of the most significant contributions in the history of biology was made by Watson and Crick in 1953 when they published a groundbreaking paper in the journal Nature. The DNA molecule's structure was identified by the two researchers as the carrier of genetic information[20]. Since then, people have realized that the four bases of DNA are designed to store the genetic information of organisms in a linear sequence Over the past decade, numerous researchers have put forth the idea of storing specific information in DNA[21], [22]. Nevertheless, the practical realization of this concept has remained elusive due to the nascent stage of DNA synthesis and sequencing technology.

The first demonstration of information storage in DNA took place in 1988[23]. 7920 bits have been coded in the largest project to date[24]. The limited scope of previous research can be attributed to the challenges associated with writing and reading lengthy, error-free DNA sequences, which restricts the potential for broader applications. In order to meet storage needs, researchers have proposed a new method of storing data. A process called genetic data storage is one of the new methods of encoding information in DNA. The main objective of this research[25] is to introduce DNA as a highly effective method for data storage and to tackle the two primary challenges associated with it. The first one is the arduous process of extracting genomic data, though advances are anticipated. The second is the cost factor, which will rise as a result of how alluring such technology might be.

In a study conducted by researchers in[17] various experiments were performed to evaluate the artificial aging of DNA. The encoded information was stored within encapsulated DNA fragments, employing error-correcting codes. To expedite the aging process, the mixture was exposed to harsh conditions, including high temperatures. By monitoring the degradation kinetics over time, the researchers successfully retrieved the original information. The findings of this experiment suggest that the artificial aging process simulates the equivalent of 2000 years in Central Europe.

The goal of this study [26] is to outline the entire process of using DNA as a data storage technology and the challenges of its adoption. This paper is special in that it examines both data storage in the DNA in vitro and data storage in DNA of living cells. In order to achieve efficient storage of DNA data, some researchers have developed efficient

and robust direct error detection and correction schemes and solutions suitable for DNA channels. In the work [27], a direct error handling strategy was proposed that can deal with all kinds of faults in the current DNA synthesis, amplification, and sequencing processes. They successfully stored and recently recovered 22 MB of digital data in one experiment using synthetic DNA. The discovered residual probability can be easily increased and is practically identical to that of the hard disk. This establishes the viability of employing artificial DNA as a long-lasting data storage medium.

3 BACKGROUND

With the development of information technology, huge amounts of data are being produced in an increasing manner. The urgent problem that exists is to find an efficient and inexpensive storage medium for this data. The genetic material DNA has attracted considerable interest as a storage medium for digital information due to its high density, durability, and persistence, making it a viable solution to future Big Data storage problems [28]–[30]. Recent studies have shown significant advances in DNA storage technology and processes [31], [32]. These advances have also led to the development of data storage technologies based on DNA molecules.

3.1 Natural DNA: structure and storage mechanism

DNA, which stands for deoxyribonucleic acid, is a molecule that stores the genetic information of an organism, i.e. all the information necessary for biological production and development [33]. The genetic code corresponds to a series of 4 nucleotides: A (adenine), C (cytosine), G (guanine) and T (thymine). Each nucleotide consists of three elements, sugar (deoxyribose), phosphoric acid and nitrogenous bases. For each combination, there will be a correspondence: each triplet corresponds to an amino acid, and the amino acids constitute the protein.

The DNA sequences in DNA are made up of four fundamental nucleotides. A DNA strand or oligonucleotide is essentially a linear (unbranched) polymer made up of a collection of nucleotides. As shown in figure 2 each nucleotide is made up of a deoxyribose sugar, which is then coupled to a phosphate group, and one of the four nitrogenous bases: adenine (A), guanine (G), thymine (T), or cytosine (C). A and T are grouped together and C and G are aligned to form a double helix when two strands of DNA are combined. These two DNA strands are hence complementary to one another [34], [35].

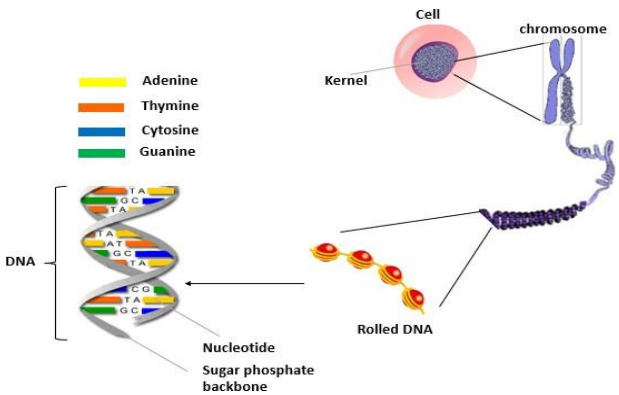

Fig. 1. DNA molecule and double Helix structure

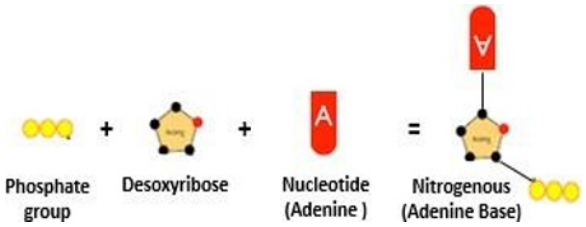

Fig. 2. Adenine Nucleotide structure

3.2 Storage process in DNA synthetic

The information in DNA is stored using four fundamental components rather than binary files, which are how computer data is often saved. These elements, denoted by the letters A, T, C, and G, are adenine, thymine, cytosine, and guanine.[36].

As a refresher, the binary system is a base-2 numbering scheme. We named it the "bit" to refer to the binary numbering's digits. A bit can have one of two values, denoted by the conventions "0" and "1". By altering the magnetic, electrical, or optical properties of the materials, traditional media (such as hard drives, USB sticks, or DVDs) retain digital data by preserving these 0s and 1s [37]. Each byte consists of eight bits. Though the method is different, the idea of storing data in DNA is the same. Digital data is typically transformed in stages before being stored in DNA to make it more suited for storage.

The data conversion process in DNA storage is different, instead of creating sequences of 0s and 1s as numeric data, DNA data storage uses sequences of nucleotides. There are several methods, but the general idea is to assign values to DNA nucleotides. This schematic storage process is shown in figure 3 includes an encoder that encodes a binary string into a DNA oligonucleotide, a DNA synthesizer that generates a string that encodes the data to be stored in the DNA, and a DNA sequencer that reads the strand and a decoder that transmits the DNA strand to the original digital data [38], [39]. This section will present and detail the basic steps of storing and retrieving digital data in DNA storage.

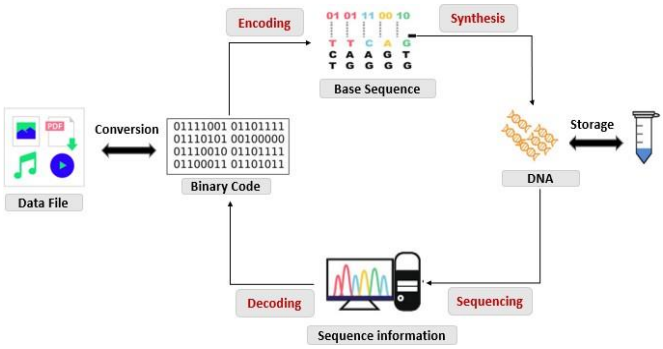

Fig.3. Overview of DNA data storage system

1) Encoding data into the DNA sequence: In order to store binary numbers in DNA, we must first convert the numerical data to DNA sequence format (i.e., A, T, G, and C). Encoding is the first step in storing DNA data. It first converts readable or visible elements such as text, numbers, and images into a binary language containing 0 and 1, and then reverses its binary information into a DNA nucleotide sequence, using four bases (A, C, G, T) in place of 0 and 1 using computer algorithms [37]. Each base pair

can be represented by 2 bits, which gives 4 different possibilities, corresponding to 16 combinations of DNA base pairs, for example (AT is 00, GC is 01, TA is 10, CG is 11) [39].

2) DNA synthesis (writing): Once the binary data is encoded and DNA sequences are generated, the subsequent step involves writing the DNA sequences into DNA molecules. This process can be achieved through chemical synthesis, as described in[39], [40]. Each nucleotide is added to its adjacent nucleotide based on the numerical data of the sequence. It is important to note that artificial DNA synthesis operates at an efficiency of 99% with a 1% error rate. The code-based protocol is specifically designed to handle text files with lines or fasta files and decode the information into stored digital files. Once the synthesis is completed, a multitude of DNA strands carrying binary data are combined in a tube or pool.

3) DNA sequencing (reading): Our assumption is that a vast number of distinct DNA strands are combined within a tube. To retrieve and restore the digital data to its original form, the complete sequencing of all the DNA strands is necessary. DNA sequencing is the process of converting DNA sequences into digital sequences. The initial step in reading the desired DNA strand involves extracting droplets from the DNA tube and amplifying the target DNA strand using PCR (Polymerase Chain Reaction). PCR is a biotechnological method that allows for the exponential replication of the target DNA sequences within a tube. During PCR, a specific primer pair is introduced to initiate the duplication of the DNA chain [36]. Presently, existing technologies enable the synthesis of DNA sequences containing up to 3,000 base pairs (bp) [41].

4) Decoding: The last stage of DNA storage is decoding. The sequence is created and delivered back to the decoder, which converts the produced sequence into binary language using a different computer program with a reverse coding feature.

3.3 Errors during DNA storage process

When storing information in DNA, a variety of errors can occur, including insertion errors, deletion errors, oligonucleotide substitution errors, and multiple oligonucleotides of different types having the same address [42].
Errors in DNA can typically be classified into three main types: substitutions, insertions, and deletions. The significance of these errors varies depending on the specific synthesis and sequencing technology employed, with most published studies indicating that substitutions or deletions tend to be the most prevalent types of errors [27], [35]. For example, the dominant errors in the on-column oligo DNA synthesis process are deletions that result either from a failure to remove dimethoxy trityl (DMT) or from combined inefficiencies in the coupling and capping steps [40].

Polymerase chain reaction (PCR) amplification cycles usually come before the sequencing step. Each DNA molecule is multiplied by around two during each cycle. As a result, the synthesis process, the PCR stages, and the DNA degradation during storage all affect how many molecules are in the pool. In conclusion, mistakes that happen during the synthesis, storage, handling, and sequencing processes include:

- ✓ There is a possibility of unsuccessful synthesis of molecules, and certain molecules may be synthesized more frequently than others. The current techniques of synthesis result in thousands or millions of copies of a DNA string, each of which may contain unique errors or variations.

- ✓ DNA degradation during storage, results in the loss of molecules.

- ✓ We can only view a subset of the molecules in the pool because the read is similar to drawing from a collection of molecules. The number of pulls (i.e., reads) we make and the distribution of the molecules in the pool both affect this score.

- ✓ DNA synthesis and sequencing can lead to the insertion, deletion, and substitution of nucleotides in individual DNA molecules.

3.4 Potential and advantages of storing data in DNA synthetic

As mentioned earlier, the exponential growth of data in the world is facing huge data storage problems. DNA has potential advantages such as high density, high replication efficiency, durability, and long-term stability, and can be used as a promising method to solve these storage problems [28].

1) Information density: Compared to the best traditional systems, DNA has an information density that is roughly 10 million times higher. One gram of DNA is thought to be capable of storing up to 215 petabytes of data (1 petabyte = 1 million gigabytes), although this estimate changes as various research teams look into new ways to test the maximum capacity of DNA storage. As a result, scientists believe that less than 100 g of DNA constitutes the current SGD of all humans. However, it is not actually a matter of synthesizing a single DNA molecule to encode a file, but of many identical copies. Moreover, areas of this DNA will have to carry quality control and indexing signals, in addition to the data. This DNA must be preserved in macroscopic containers [43].

2) Longevity: DNA has a lifespan that is roughly ten thousand times longer than that of conventional media. From historical samples, DNA molecules older than 560,000 years have been analysed. It's artificially accelerated aging in the lab has demonstrated that its half-life is 52,000 years [44]. Additionally, because DNA is stable at room temperature, its preservation is unaffected by cold, making this storage method not especially energy-intensive.

3) Calculations: The physic-chemical properties of DNA facilitate the direct execution of certain calculations. The principle of this calculation is to use synthetic DNA strands to code a combined problem, manipulate these strands with molecular biology tools to simulate the functioning of the separation solution, and then read the latter by sequencing [45].

4) Reduced carbon footprint: Synthetic DNA enables large amounts of data to be stored in small spaces, reducing the need for traditional storage infrastructure. This in turn helps to reduce the energy consumption needed to power and cool these infrastructures, thereby reducing greenhouse gas emissions and the overall carbon footprint.

3.5 challenges of DNA storage technology

DNA may become an attractive medium for digital data storage due to its distinctive qualities when compared to existing media. However, before DNA can be used

commercially, there is still much work to be done. Short synthetic DNA fragments, high synthesis and sequencing error rates, low throughput, limited data storage access, and high prices are just a few of the difficulties we face.

Currently, the cost of DNA synthesis and sequencing still exceeds any feasible real-world application. Over the past decade, the cost of DNA synthesis and sequencing has dropped by several orders of magnitude. However, due to the artificial influence of market prices, the rate of cost decline has slowed in recent years.

4 EXPERIMENT AND DISCUSSION

The goal of this section is to present an experiment comparing the performance of two error detection codes used in DNA storage—the Hamming code and the Cyclic Redundancy Check (CRC) code. The experiment aims to encode files stored in memory as DNA sequences and decode those same DNA sequences back to their original format. In order to prevent mutations and data damage, we alternately utilize the Hamming code and the CRC code, and we compare the performances of these two codes simultaneously.

The program can read any ASCII text file and encode it into a DNA sequence aligned to the 13 data bits of the Hamming (12, 8) code. In other words, there are 5 parity bits for 8 bits of raw data. The output is in FASTA format. In order to restore its original content, the same procedure can be used to decode the encoded DNA sequence, with a file encoded in FASTA format. Repeat the same process using a cyclic redundancy checking (CRC) code to compare its performance and efficiency with the Hamming code.

4.1 Hamming Code

The future development of computers requires greater reliability, in particular the ability to detect and correct errors. This demand prompted Richard Hamming to first introduce error correction and detection codes in 1995. The initial work on Hamming's code enabled large computers to perform a large number of operations, and there was not a single error in the final result. Hamming code ensures that the transmitted/stored information will not be damaged or affected by a single-bit error. This type of binary code typically uses K parity bits added to n data bits to form a new word of (n + k) bits [46].

The Hamming code is a binary code that consists of data bits and parity bits every 2n positions. The checksum function of various subsets of data bits uses the parity bit, which enables the detection of replacement errors [42]. Hamming employs a relatively complex checksum algorithm; the position 1 of the first parity bit serves as a parity check for the place where the least significant bit in the binary representation is 1. (i.e. positions 1, 3, 5, 7...). Position 2 contains the second check bit, which is a parity check with position 1 being the second least significant bit in its binary representation (i.e. positions 2, 3, 6, 7,...). If no parity fails, the code word is assumed to be correct. Furthermore, if a certain bit in the codeword is misrepresented, the error is where the binary representation equals the failed parity mode [46].

4.2 Cyclic redundancy checking (CRC) code

A crucial kind of error detection code is the cyclic redundancy check (CRC), which is a widely accepted industry standard for error detection in data transfers. Because they are quick and simple to implement in encoders and decoders and have strong burst error detection capabilities, these codes are commonly employed in computer communication networks. The abbreviated cyclic code structure offers these features. The capacity to detect burst errors has been thoroughly researched in [47]. Despite the existence of more efficient

data correction codes, CRC continues to be widely utilized in embedded communication protocols due to its simplicity, minimal code, and hardware overhead. [48]. The cyclic redundancy check code can be considered as a polynomial code, and the transmitted bit string can be interpreted as a polynomial whose coefficients are the 0 and 1 values of the bit string because of each codeword C(x)=Cn-1Cn- 2... C0 can be expressed by a polynomial of degree n-1, as shown by the following equation.

$$C(x) = \sum n-1 \quad Ci\ Xi$$

4.3 Result and discussion

Figures 4 and 5 illustrate the experimental results of converting a line-format text file to DNA sets. The program takes the input text file as the first parameter and calls the encoding function used to perform the conversion; the result is stored in a file in Fasta format. The conversion is performed the first time using the Hamming code, and the second time using the Cyclic redundancy checking (CRC) code to compare their performance.

The FASTA format, also known as the Pearson format, is a file format employed for storing biological sequences related to nucleic acids or protein properties. These sequences are represented by a series of letters that encode nucleic acids or amino acids. Each sequence can be accompanied by a name and a comment. Files in the FASTA format are typically represented by .fasta or .fa extensions. The simplicity of the FASTA format makes it easy to manipulate and read (or analyze) sequences using word processing tools and scripting languages (such as Python, R, Ruby or Perl).

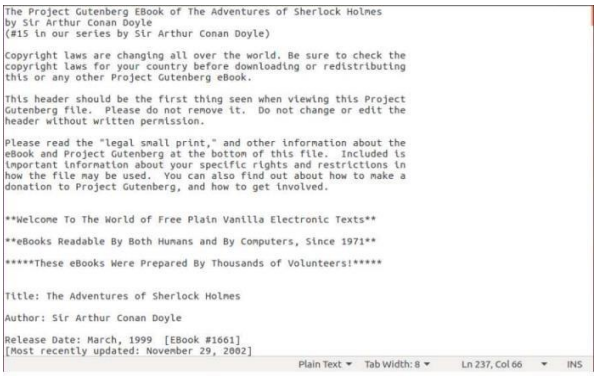

Fig.4. The text file

Fig.5.The output Fasta format file

The decoding function is used to convert the 12-bit aligned FASTA DNA file generated by the encoding operation into its original error-free form. The result is the original text file.

Table 1. Data rate comparison between Hamming and CRC codes

Data transmitted (bytes)	**Data Rate (frames per second)**	
	Hamming	**CRC**
1	15,452	14,070
2	14,046	12,780
3	12,781	11,580
4	12,562	10,403
5	10,389	9,273
6	9,245	8,152
7	8,136	7,060
9	7,130	6.032
16	6,116	6.005
32	4.100	2.620

The comparison of the simulated transmission rates for CRC and Hamming codes in a 32-bit system is shown in Table 1. The effectiveness of the two codes was examined during the experience to see which one would provide a better data rate for error detection and correction. It demonstrates that the Hamming code has a quicker transmission rate than CRC codes.

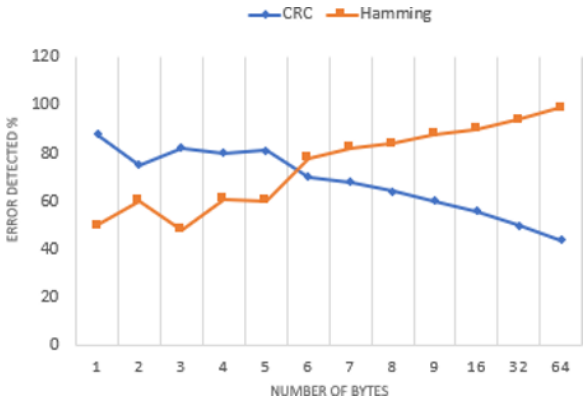

Fig.6. Error detection performance of CRC and Hamming Code

CRC is an error detection method with numerous algebraic properties that facilitate the implementation of encoding and decoding. Unlike the Hamming code, which adds additional binary bits to guarantee the integrity of the original code, CRC is designed based on the remainder obtained from polynomial division. While CRC can be time-consuming for cases involving single errors, it is more efficient to utilize CRC in combination with Hamming codes specifically for detecting burst errors. This approach allows for a

simplified implementation of encoding and decoding while maintaining effective error detection capabilities.

As shown in Figure 6, when the number of bits is small, the error detection performance of CRC is better than that of Hamming code, but when the length of the code word increases (the number of bits increases), the error detection performance of Hamming code far exceeds that of CRC code. Therefore, in our DNA storage case, we need a code that can reliably perform error detection on a large number of bits. Ignoring the error detection performance of the CRC code with a word length of some bits, we can conclude that as the code word length increases, the Hamming code can detect a larger percentage of errors than CRC.

5 CONCLUSION AND FUTURE WORK

With the advent of the era of big data, mobile applications, social media, and megadata analytics plans, the world has experienced explosive data growth and the challenge of big data storage is growing. The right storage device must be chosen in order to calculate the growth rate appropriately. When choosing the best storage solution system, factors like capacity, performance, throughput, cost, scalability, and reliability are crucial. The methods employed for storing and managing Big Data can have a profound impact on the overall functioning of an organization. It is crucial to prioritize solutions that address the present and future challenges of Big Data storage, focusing on identifying more efficient alternatives to meet evolving requirements.

With this explosion in the amount of data, natural storage seems to be the solution to preserve data as an archive for a longer period of time. DNA for data storage is a persistent information storage solution that allows us to store large amounts of data in a very small space. Due to its extremely high density and long shelf life, DNA data storage has become one of the most cutting-edge technologies for long-term data storage. Data is compressed and data security is ensured. Unlike traditional storage which imposes restrictions on the form of the encoded data. The data stored in the DNA storage system is also subject to two biochemical constraints.

One of the major problems of DNA storage was that the error rate in the coding of information was naturally high. To solve the problem of errors in the chemical storage of DNA and to achieve a high storage density with reliable retrieval of the archived data, researchers use several solutions to detect and correct these errors.

In this paper, we try to use two error detection and correction codes, the CRC code and Hamming code, as error detection and correction techniques. So, the performance of both codes has been compared. As the results show, the use of Hamming code allows to obtain a faster data rate, its performance at the level of error detection is important in front of the CRC code when the number of bits is high. In addition, other improvements can be obtained by developing error detection and correction codes in order to improve the limits of the use of cyclic redundancy check (CRC) codes and Hamming codes. For example, these two codes can be grouped together in order to recombine their performances and have more practical solutions.

This work was supported in part by the National Center for Scientific and Technological Research (CNRST) and this within the program of the research grants initiated by the Ministry of National Education, Higher Education, Management Training and Scientific Research.

References

[1] V. Belov, A. Tatarintsev, and E. Nikulchev, "Comparative Characteristics of Big

Data Storage Formats," Journal of Physics Conference Series, vol. 1727, p. 012005, Jan. 2021, doi: 10.1088/1742-6596/1727/1/012005.

[2] A. Gusev, D. Ilin, and E. Nikulchev, "The Dataset of the Experimental Evaluation of Software Components for Application Design Selection Directed by the Artificial Bee Colony Algorithm," Data, vol. 5, no. 3, Art. no. 3, Sep. 2020, doi: 10.3390/data5030059.

[3] G. Petushkov, "Evaluation and reliability prediction for highly reliable software and hardware systems: The case of data processing centers," Russian Technological Journal, vol. 8, pp. 21–26, Mar. 2020, doi: 10.32362/2500-316X-2020-8-1-21-26.

[4] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," Mob. Netw. Appl., vol. 19, no. 2, pp. 171–209, Apr. 2014, doi: 10.1007/s11036-013-0489-0.

[5] P. Russom, "Big data analytics," TDWI best practices report, fourth quarter, vol. 19, no. 4, pp. 1–34, 2011.

[6] "Cloudera administration handbook." https://text.123docz.net/document/5338213-cloudera-administration-handbook-rohit-menon-5-pdf.htm (accessed Oct. 21, 2021).

[7] "HortonWorks Data Platform : new book," 2015.

[8] T. Dunning and E. Friedman, Real-World Hadoop. O'Reilly Media, Inc., 2015.

[9] D. Quintero et al., Implementing an IBM InfoSphere BigInsights Cluster using Linux on Power, First edition. in IBM redbooks. IBM, International Technical Support Organization, 2015.

[10] "Pivotal HD Enterprise 1.1 Installation and Administrator | Manualzz," manualzz.com. https://manualzz.com/doc/25974984/pivotal-hd-enterprise-1.1 installation-and-administrator (accessed Aug. 07, 2022).

[11] D. Sarkar, "Pro Microsoft HDInsight : Hadoop on Windows /," 2014.

[12] J. Moorthy et al., "Big Data: Prospects and Challenges," Vikalpa:The Journal for Decision Makers, vol. 40, pp. 74–96, Mar. 2015, doi: 10.1177/0256090915575450.

[13] Z. Ping et al., "Carbon-based archiving: current progress and future prospects of DNA-based data storage," GigaScience, vol. 8, no. 6, p. giz075, Jun. 2019, doi: 10.1093/gigascience/giz075.

[14] M. Castillo, "From Hard Drives to Flash Drives to DNA Drives," AJNR. American journal of neuroradiology, vol. 35, Apr. 2013, doi: 10.3174/ajnr.A3482.

[15] R. Appuswamy et al., "OligoArchive: Using DNA in the DBMS storage hierarchy," in CIDR, 2019.

[16] M. E. Allentoft et al., "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils," Proceedings of the Royal Society B: Biological Sciences, vol. 279, no. 1748, pp. 4724–4733, Dec. 2012, doi: 10.1098/rspb.2012.1745.

[17] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," Angew Chem Int Ed Engl, vol. 54, no. 8, pp. 2552–2555, Feb. 2015, doi: 10.1002/anie.201411378.

[18] R. Heckel, G. Mikutis, and R. N. Grass, "A Characterization of the DNA Data Storage Channel," Sci Rep, vol. 9, no. 1, p. 9663, Dec. 2019, doi: 10.1038/s41598-019-45832-6.

[19] C. Roy, M. Pandey, and S. SwarupRautaray, "A Proposal for Optimization of Data Node by Horizontal Scaling of Name Node Using Big Data Tools," in 2018 3rd International Conference for Convergence in Technology (I2CT), Pune: IEEE, Apr.

2018, pp. 1–6. doi: 10.1109/I2CT.2018.8529795.

[20] J. D. Watson and F. H. C. Crick, "THE CLASSIC: Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," Clinical Orthopaedics and Related Research®, vol. 462, pp. 3–5, Sep. 2007, doi: 10.1097/BLO.0b013e31814b9304.

[21] M. Neiman, "Some fundamental issues of microminiaturization," Radiotekhnika, vol. 1, pp. 3–12, 1964.

[22] M. Neiman, "On the molecular memory systems and the directed mutations," Radiotekhnika, vol. 6, pp. 1–8, 1965.

[23] J. Davis, "Microvenus," Art Journal, vol. 55, no. 1, pp. 70–74, Mar. 1996, doi: 10.1080/00043249.1996.10791743.

[24] D. G. Gibson et al., "Creation of a bacterial cell controlled by a chemically synthesized genome," Science, vol. 329, no. 5987, pp. 52–56, Jul. 2010, doi: 10.1126/science.1190719.

[25] J. P. Cox, "Long-term data storage in DNA," Trends Biotechnol, vol. 19, no. 7, pp. 247–250, Jul. 2001, doi: 10.1016/s0167-7799(01)01671-7.

[26] L. Ceze, J. Nivala, and K. Strauss, "Molecular digital data storage using DNA," Nat Rev Genet, vol. 20, no. 8, pp. 456–466, Aug. 2019, doi: 10.1038/s41576-019-0125-3.

[27] M. Blawat et al., "Forward Error Correction for DNA Data Storage," Procedia Computer Science, vol. 80, pp. 1011–1022, Jan. 2016, doi: 10.1016/j.procs.2016.05.398.

[28] G. Church, Y. Gao, and S. Kosuri, "Next-Generation Digital Information Storage in DNA," Science (New York, N.Y.), vol. 337, p. 1628, Aug. 2012, doi: 10.1126/science.1226355.

[29] N. R. MANAR SAIS JAAFAR ABOUCHABAKA, "SYNTHETIC DNA AS A SOLUTION TO THE BIG DATA STORAGE PROBLEM," Journal of Theoretical and Applied Information Technology, vol. 99, no. 15, Aug. 2021, doi: 10.5281/zenodo.5353710.

[30] M. Sais, N. Rafalia, and J. Abouchabaka, "Intelligent Approaches to Optimizing Big Data Storage and Management: REHDFS system and DNA Storage," Procedia Computer Science, vol. 201, pp. 746–751, Jan. 2022, doi: 10.1016/j.procs.2022.03.101.

[31] M. T. Barrett et al., "Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA," Proc Natl Acad Sci U S A, vol. 101, no. 51, pp. 17765–17770, Dec. 2004, doi: 10.1073/pnas.0407979101.

[32] Z. Chen et al., "Highly accurate fluorogenic DNA sequencing with information theory–based error correction," Nat Biotechnol, vol. 35, no. 12, Art. no. 12, Dec. 2017, doi: 10.1038/nbt.3982.

[33] D. Limbachiya and M. Gupta, "Natural Data Storage: A Review on sending Information from now to then via Nature," May 2015.

[34] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "The Structure and Function of DNA," Molecular Biology of the Cell. 4th edition, 2002, Accessed: Jul. 08, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK26821/

[35] Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," Synthetic Biology, preprint, Sep. 2016. doi: 10.1101/074237.

[36] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in DNA

by primed synthesis with DNA polymerase," J Mol Biol, vol. 94, no. 3, pp. 441–448, May 1975, doi: 10.1016/0022-2836(75)90213-2.

[37] B. Li, L. Ou, and D. Du, "IMG-DNA: Approximate DNA Storage for Images," Mar. 2021, doi: 10.1145/3456727.3463771.

[38] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding Over Sets for DNA Storage," IEEE Trans. Inform. Theory, vol. 66, no. 4, pp. 2331–2351, Apr. 2020, doi: 10.1109/TIT.2019.2961265.

[39] C. Bancroft, T. Bowler, B. T, and C. T. Clelland, "Long-term storage of information in DNA," Science, vol. 293, no. 5536, pp. 1763–1765, Sep. 2001, doi: 10.1126/science.293.5536.1763c.

[40] S. Kosuri and Church, "Large-scale de novo DNA synthesis: technologies and applications," Nat Methods, vol. 11, no. 5, pp. 499–507, May 2014, doi: 10.1038/nmeth.2918.

[41] S. M. H. Tabatabaei Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A Rewritable, Random-Access DNA-Based Storage System," Sci Rep, vol. 5, no. 1, p. 14138, Sep. 2015, doi: 10.1038/srep14138.

[42] R. W. Hamming, "Error Detecting and Error Correcting Codes," Bell System Technical Journal, vol. 29, no. 2, pp. 147–160, 1950, doi: 10.1002/j.1538-7305.1950.tb00463.x.

[43] L. Organick et al., "Random access in large-scale DNA data storage," Nat Biotechnol, vol. 36, no. 3, pp. 242–248, Mar. 2018, doi: 10.1038/nbt.4079.

[44] J. Bonnet et al., "Chain and conformation stability of solid-state DNA: implications for room temperature storage," Nucleic Acids Res, vol. 38, no. 5, pp. 1531–1546, Mar. 2010, doi: 10.1093/nar/gkp1060.

[45] L. M. Adleman, "Molecular computation of solutions to combinatorial problems," Science, vol. 266, no. 5187, pp. 1021–1024, Nov. 1994, doi: 10.1126/science.7973651.

[46] A. Ahmadpour and A. Ahadpour Shal, A Novel Formulation of Hamming Code. 2009. doi: 10.1109/ECTICON.2009.5137169.

[47] S. B. Wicker, Error Control Systems for Digital Communication and Storage, US e. édition. Englewood Cliffs, NJ: Pearson, 1994.

[48] M. W. Azhar, T. T. Hoang, and P. Larsson-Edefors, "Cyclic Redundancy Checking (CRC) Accelerator for the FlexCore Processor," in 2010 13th Euromicro Conference on Digital System Design: Architectures, Methods and Tools, Sep. 2010, pp. 675–680. doi: 10.1109/DSD.2010.51.