Towards a Robust DNA Storage System with a Multilayer Approach

Ayoub Sghir, Manar Sais, Douha Bourached, Jaafar Abouchabaka, Najat Rafalia
Department of Computer Science, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

***Abstract*—The growing demand for data storage requires innovative, resilient solutions to the challenges of cost, space, and energy consumption posed by current methods. DNA stands out as a promising next-generation data storage medium, offering a remarkable storage density of 1019 bits per cubic centimeter, some eight orders of magnitude denser than conventional media. This study explores the potential of DNA storage by proposing an intelligent multi-layer solution to overcome current technological challenges. The system combines the storage capabilities of DNA with sophisticated solutions such as data compression, error correction, and cryptography, transforming the concept of DNA storage into a tangible reality. This study also focused on the first layer dedicated to data compression. The results obtained represent a significant advance in the evaluation of the potential of different compression algorithms, through a comparative study of techniques such as Huffman coding, run-length coding, LZW and LZ77. This analysis enabled us to define the essential components of the first layer of the proposed approach. Finally, the interface in the digital domain to visually present the overall results of the project was introduced, while providing insight into the system's efficiency, data integrity and ease of use.**

***Keywords*—DNA storage; data compression; error correction; huffman encoding; run-length encoding; LZW; LZ77**

I. Introduction

The massive growth in data over the last few decades, particularly with the development of Internet and mobile technologies, has led to an unprecedented data storage crisis for mankind. In social and scientific research, as well as in other fields such as traffic monitoring and high-definition scientific images, massive data, particularly in the form of images and videos, is generated. According to international data [1], the total amount of human-generated data reached 33 ZB (zettabytes) in 2018, with a forecast of 175 ZB by 2025, a 65-fold increase in just thirteen years [2], [3], [4]. Traditional data storage methods such as magnetic, optical and solid-state disks have major shortcomings in terms of cost, space, energy consumption and durability, not exceeding 50 years under optimal conditions.

In the age of Big Data, the need for innovative, sustainable storage solutions is becoming ever more pressing. DNA [5] is emerging as a promising candidate for long-term storage, thanks to its exceptional durability and extraordinary information density, offering a future where vast data sets can be encoded in the very fabric of life. However, to fully exploit the potential of DNA storage [6], technological challenges such as security, reliability and turnaround time must be carefully overcome. The ability to safeguard data, guarantee its integrity and speed up recovery is crucial to the success and widespread adoption of DNA storage.

DNA storage has a rich history dating back to 1959, when Richard Feynman first proposed the concept of storing information on DNA. Since then, interest in this technology has grown considerably due to the many potential advantages it offers over traditional storage methods such as magnetic and optical storage. DNA stands out for its exceptional stability, high storage density and long-term durability, making it a promising solution to the data storage challenges faced by traditional media [7-9]. Recent advances have demonstrated that DNA can efficiently store vast quantities of data, including historical audio and video files, with remarkable density and durability, positioning it as a suitable medium for digital archives [10]. Furthermore, according to research proposals, it could transform the landscape of data storage for network automation and prediction. This integration would offer an efficient and reliable way of storing large amounts of data generated in network operations, paving the way for new possibilities in large-scale data optimization and management.DNA storage has attracted considerable interest due to its high information density and longevity. Various approaches have been proposed to optimize DNA storage methods.

In this paper [11], the authors suggest using DNA as a storage approach, demonstrating that DNA has both extremely high theoretical data density and unrivalled stability compared with more traditional data storage media. In another work [12] A new random sampling channel model, built around the three essential dimensions of DNA storage systems: 1- information is stored on several DNA molecules; 2- these molecules undergo noise perturbations during synthesis and sequencing; and 3- data is extracted by random sampling from the DNA reservoir.[13] suggested a powerful, feasible and particularly robust coding algorithm called MOPE, which was used to develop the no-cost coding system. Payload encoding was also carried out using the Payload Encoding algorithm [14]. In another study [15], the authors analyze the impact of batch optimization on reducing the cost of large-scale DNA synthesis. This optimization consists of the following algorithmic task: from a large set of random quaternary chains of specified length, divide S into groups in order to minimize the sum of the lengths of the shortest common super sequences in these groups.

Despite significant progress in the field of DNA data storage, no truly efficient system is yet available to manage the entire storage process in an optimal way. Numerous challenges remain with regard to the reliability, efficiency and safety of this storage method. Indeed, the biological aspect of DNA gives rise to specific difficulties, such as faults occurring during the synthesis, amplification and sequencing processes, which can impact on the quality of recorded data. What's more, the absence of a unified system for compression, error

correction and encryption significantly reduces the potential of this promising technology. This problem highlights the need to design an innovative and efficient system capable of overcoming these challenges and offering an integral solution for managing the DNA data storage process.

This study aims to embark on a journey to meet the challenges head-on. By introducing a revolutionary multi-layer approach to DNA storage management, the study aims not only to unleash its full potential but also to usher in a new era of secure, reliable, and efficient data storage. In the following pages, we'll look at the intricacies of this approach, which combines data compression, error correction, and encryption, each playing a key role in reshaping the DNA storage landscape. The careful selection of data compression algorithms is revealed [16]. The introduction of polar codes as an error correction mechanism will provide the basis for data reliability, considering the defects inherent in DNA synthesis, amplification and sequencing. To improve the security of DNA storage, we are introducing PGP encryption, a widely recognized encryption standard, by adapting it to the unique properties of DNA. In doing so, we bring to DNA storage the same level of data security found in the world of encrypted hard disks, opening up a new era of trust in data storage solutions.

The implementation of these layers demonstrates a commitment to advancing DNA storage technology. By converting binary data into ACTG sequences and vice versa, the study ensures seamless integration of this multi-layer approach into the DNA storage ecosystem. By exploring the intricacies of this study, the aim is to highlight the remarkable synergy between biology and information technology, illustrating how DNA's exchange capabilities can transcend the limits of traditional data storage methods.

Through this holistic approach, the work aims to reduce turnaround times, improve data reliability and enhance the security of DNA storage, by presenting it as a secure, high-performance option for the long-term maintenance of large volumes of data. This multi-layer approach aims to guarantee a reliable data storage process in DNA, focusing on the first and essential layer: data compression. This experimental section focuses on this compression layer, examining and comparing different compression techniques such as Huffman coding [17], run-length coding (RLE) [18], LZW [19] and LZ77 [20]. This analysis enabled us to identify the most effective algorithm for the developed approach. The results show that, although some of these algorithms can be adapted to other types of data, the Huffman algorithm proves to be the most reliable for this approach and also for this type of data, Thanks to its ability to efficiently manage the frequency of symbols, it also provides a user interface (UI) in the digital domain, enabling visualization of global results related to the compression layer. This interface provides insight into the system's efficiency, data integrity and ease of use, contributing to a better understanding and exploitation of the performance of this approach.

The structure of the following study is as follows: Section II explores the framework of DNA storage processes, Section III details the presented method and system phases, Section IV discusses testing and algorithm evaluation, while Section V presents the conclusions.

II. Background

Not too long ago, the terms 'Big' and 'Data' were rarely associated. Today, they have become a highly fashionable expression and rank among the most popular concepts in the business domain. Numerous authors and organizations have made attempts to define the term Big Data. Hemn Barzan Abdalla wrote: "Big Data is described as a high-throughput data resource that requires new processing measures to achieve better knowledge dynamics and disclosure" [21]. He further elucidated "too voluminous" about the massive volume of data that can reach the scale of petabytes and originates from diverse sources, "too rapid" concerning the rapid growth of data that must be processed swiftly, and "too complex" to denote the challenges of Big Data that do not readily integrate with existing processing tools. Similarly, in PCMag, one of the most popular publications on technology trends, Big Data is defined as 'enormous quantities of data accumulated over time, which are challenging to analyze and manage using conventional database management tools' [22].

Currently, the existing data volume is measured in petabytes and is projected to increase to the order of zettabytes shortly. Even today, established social media platforms generate terabytes of data daily, a quantity that undoubtedly poses challenges for traditional systems [23]. According to statistics, the daily data generation amounts to approximately 44 zettabytes. Each second, an individual generates 1.7 megabytes of data [24]. According to forecasts from the International Data Group, global data quantities are set to experience exponential growth from 2020 to 2025, with an expected increase from 44 to 163 zettabytes [25]. Consequently, the significance of storage has never been greater when it comes to managing the ever-expanding data volumes.

The primary obstacle posed by the realm of big data revolves around identifying a technology capable of efficiently handling substantial data quantities for subsequent analysis and data extraction. Numerous solution providers offer ready-made answers to confront the challenges of big data, including Cloudera [26], Horton Works[27], MapR [28], and more. However, the question of real-time management in terms of temporality and storage space remains one of the major challenges of contemporary technologies. However, the explosive growth in data volumes challenges these existing data storage and processing capabilities, driving the need for innovative and unconventional solutions.

The concept of DNA storage, which represents a significant breakthrough in the data storage sector, came about thanks to the introduction of a unified system combining compression, error correction, and encryption. This multi-level method facilitates the management of the DNA data storage process, while ensuring both reliability and compression.

A. Leveraging the Data Storage Capacity of DNA

As our world generates massive amounts of data, conventional storage methods, such as hard disks and magnetic tapes, are now reaching their limits in terms of density and longevity. However, new molecular media, such as small organic molecules [31], [32], polymers[33] and especially DNA, are attracting growing interest for their intrinsic storage capacity, High-density data. DNA could be a perfect option

for archiving digital data, with notable advances in coding schemes enabling a revolutionary density of 17 exabytes per gram, far surpassing magnetic and optical media [34].

Since the beginning of existence, the planet's nature has created its own data storage mechanism by incorporating the genetic information of living organisms into DNA, a molecule containing four different bases: adenine (A), thymine (T), cytosine (C) and guanine (G). This method of preserving information has stood the test of time for three billion years. In the context of massive data flow production, digital storage systems present themselves as a crucial and effective option, especially in terms of cost-effectiveness. What's more, DNA offers considerable advantages such as high data density, durability and long-term stability. It represents a promising response to the challenges posed by the continuing rise of massive data.

DNA has become an attractive choice in response to the growing demand for data storage, offering astonishing storage capacity [35]. Virtually all data on the Internet can be encapsulated, and in particular, it has an extraordinarily High storage capacity, indicating that a small fragment of DNA can contain a vast amount of information [36]. DNA can theoretically achieve an impressive storage density of 455 exabytes per gram. with the exceptional feature of longevity capable of preserving data for several centuries [37]. In ecological terms, DNA synthesis can be carried out in an environmentally friendly way, avoiding and minimizing the use of harmful chemicals while adopting sustainable practices. By developing environmentally responsible techniques for DNA production, storing information in synthetic DNA could promote more sustainable and responsible management of this limited available resources.

B. DNA Storage Process

DNA storage is a revolutionary advancement that has garnered significant attention in recent years. By harnessing the unique properties of DNA, including its high storage capacity and exceptional durability, researchers have developed innovative techniques for encoding and storing large amounts of digital information. This process involves converting digital data into sequences of DNA molecules, which are then synthesized and compactly stored. This approach holds the promise of revolutionizing data storage by enabling the preservation of vast amounts of information in a minuscule space.

DNA-based information processing represents an emerging field that leverages the power of DNA molecules to perform various computational tasks. These tasks include storage, addressing, transportation, and error correction of information [38]. This burgeoning field offers considerable potential to transform current computing capabilities. The structure of DNA itself is crucial for understanding various biological processes [39]. It provides a mechanism of heredity, where genes carry biological information that must be accurately copied and passed on to the next generation. Furthermore, the structure of DNA has enabled answers to fundamental questions in biology about the storage and transmission of genetic information. This in-depth understanding of DNA structure opens up new avenues for scientific research and technological advancements.

The entire DNA storage process involves six key steps: encoding, synthesis, storage, retrieval, sequencing, and decoding. Encoding translates binary data into A, C, G, T sequences, with specific encoding and error correction required. Synthesis transforms these sequences into real DNA molecules, typically generating a pool of identical molecules. The DNA molecules are then stored in small containers suitable for long-term preservation. Retrieval involves selecting desired DNA molecules from a container, often using PCR based on sequence indexing. Sequencing translates DNA molecules back into sequences of A, C, G, T using DNA sequencers. Finally, decoding corrects and reorganizes the sequences to reconstruct the original document. The synthesis stage is a critical bottleneck due to its slow speed and high cost, making DNA storage less economically viable. In contrast, sequencing is not a bottleneck, benefiting from significant advancements in speed and efficiency driven by genomic needs [40].

C. Potential and Advantages of DNA Storage

As previously mentioned, the exponential growth of global data presents significant storage challenges. DNA emerges as a promising solution due to its advantages, including remarkable information density, durability, and long-term stability [41].

1) Density: DNA's information density surpasses traditional systems by approximately ten million times. A single gram of DNA is estimated to store up to 215 petabytes of data, although ongoing research may refine this estimate. However, DNA storage doesn't involve a single molecule encoding a file; it requires numerous identical copies. Additionally, DNA segments must accommodate quality control and indexing signals alongside the data, necessitating macroscopic containers.

2) Longevity: DNA's lifespan exceeds that of conventional media by roughly ten thousand times. Even DNA molecules over 560,000 years old from historical samples can be analyzed. Laboratory experiments have demonstrated a half-life of 52,000 years for artificially aged DNA, and its stability at room temperature makes it an energy-efficient preservation method. Reduced Carbon Footprint: Synthetic DNA cleverly stores large quantities of information in small spaces, enabling optimal use and management of the storage structure, thus reducing the need for conventional infrastructure. This reduces the energy consumption required to operate and cool these facilities, leading to lower greenhouse gas emissions and a reduced overall carbon footprint.

III. Methods and Implementation

In this section, an outline of the main components and layers of the proposed multi-layer system for DNA storage management. The methodology encompasses the key components of this approach, including data compression, error correction, and encryption, as well as the conversion between binary data and ACTG sequences. The proposed multi-layer system integrates data compression and error correction, and encryption techniques to enhance the efficiency, reliability, and security of DNA storage. The process can be visualized as a well-orchestrated sequence of steps, with each contributing to the successful encoding and preservation of data in DNA form. Fig. 1 outlines the sequential steps involved in implementing this approach, beginning with data compression

and proceeding through error correction, encryption, and data format conversion. The following subsections will detail each step of the implementation process, providing a comprehensive understanding of the challenges of DNA storage.

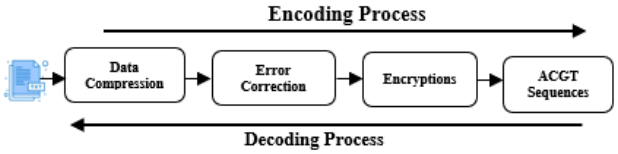

Fig. 1. Optimization storage techniques.

A. Data Compression

Data compression is the first layer of this approach, aimed at minimizing the size of textual data, thereby reducing the time required for storage and retrieval. Initially, the study considered three compression algorithms, Huffman, RLE, and LZW, to assess their suitability for this system. The primary objective of this layer is to make the data more compact and storage-friendly, setting the stage for efficient DNA-based storage. In the testing phase, the focus will be specifically on this layer, identifying the most efficient and suitable algorithm for our use case by comparing four algorithms. This step is critical to ensure that the chosen compression method aligns with the requirements of DNA storage, balancing compression efficiency, computational complexity, and compatibility with downstream processes.

1) Huffman encoding: Huffman encoding is a variable-length coding technique that efficiently represents characters by developing more concise codes for commonly used characters, and more extensive codes for less common characters. This method, effective for compressing textual data, involves creating a Huffman tree, where characters are encoded as binary codes. Implemented as part of a data compression layer, it optimizes character representation to minimize the amount of DNA required for storage before encoding data into DNA [42].

2) RLE encoding: Run-Length Encoding (RLE) is a data compression technique that simplifies textual data by representing consecutive repeated characters with a count of how many times a character is repeated, followed by the character itself. RLE proves particularly effective for highly repetitive data, efficiently reducing data size by encoding repeating sequences with concise representations, such as encoding "AAAAABBBCCDAA" as "5A3B2C1D2A". Initially considered for this multi-layer system for DNA storage management, RLE was surpassed by Huffman encoding due to its superior performance in reducing data size while maintaining integrity. While RLE offers simplicity and ease of implementation, the decision to prioritize Huffman encoding was based on its overall effectiveness, with comparative results between the two to be presented in the results chapter for comprehensive assessment within the context of DNA storage [18].

3) LZW encoding: LZW encoding, also known as Lempel-Ziv-Welch encoding, stands as a prominent data compression technique with broad applications spanning bioinformatics and communication systems. At its core, the LZW algorithm employs a dictionary-based approach, systematically replacing patterns within the data with codes to optimize both storage and transmission efficiency. Over time, various iterations and enhancements of LZW have emerged, aimed at refining compression ratios and expediting encoding processes to meet evolving demands across diverse domains [29], which optimizes storage space and the bandwidth required for transmission [8].

4) LZ77 encoding: LZ77 encoding, a widely adopted lossless data compression algorithm, operates by segmenting input strings into phrases through the analysis of previous occurrences. This method leverages a hashing table to efficiently pinpoint recurring sequences by cross-referencing data within the stream, facilitating effective compression. The algorithm analyzes the input data stream by searching for repeating sequences and breaks it into phrases consisting of both literals (individual characters) and pointers to previous occurrences of substrings within the stream. These pointers typically consist of two parts: a distance and a length, indicating how far back in the stream the substring occurs and how long it is, respectively [20].

B. Error Correction

In the second layer of this approach, attention is given to the crucial aspect of error correction in DNA storage, recognizing the potential for errors during synthesis, amplification, and sequencing processes. To ensure data integrity, this study implements Polar codes, renowned for their robust error correction capabilities. Polar codes, a class of error-correcting codes, achieve effectiveness by polarizing bits in a data stream through a mathematical transformation. This polarization distinguishes some bits as highly reliable and others as less reliable, introducing redundancy to facilitate reliable transmission. The encoding process selectively polarizes specific bits, creating encoded data for more robust transmission or storage. Upon reception, the inverse transformation is applied to decode the data and rectify errors. In the presented multi-layer system, Polar codes play a pivotal role in addressing errors inherent in DNA storage processes, significantly enhancing data reliability and reinforcing DNA's potential as a long-term storage medium [30].

C. Data Encryption

The security of data stored in DNA is of paramount importance, especially for sensitive information. To ensure confidentiality and integrity, the presented multi-layer system employs encryption techniques, specifically Pretty Good Privacy (PGP) encryption. PGP, a widely used method, combines symmetric and asymmetric key algorithms to safeguard digital data. It involves key generation, where a public key encrypts the data and a private key decrypts it. This process secures the data during storage and retrieval. PGP encryption provides robust security for DNA-stored data, ensuring protection against unauthorized access even if the DNA is compromised. By integrating PGP encryption into the developed system, an approach addressing the critical concern of data security is presented, guaranteeing confidentiality and complementing the data compression and error correction layers for a comprehensive DNA storage solution.

D. Data Format Conversion ACGT Sequence

The final layer in the multi-layer system presented in this study, the DNA storage system focuses on converting binary data into sequences of Adenine (A), Thymine (T), Cytosine (C), and Guanine (G), collectively known as ACTG sequences, which are essential for efficient encoding and retrieval of data in DNA molecules. This conversion process maps each binary digit (0 or 1) to its corresponding nucleotide base in DNA, bridging the digital and biological realms. During encoding, binary data is systematically translated into ACTG sequences to ensure accurate representation of each digit. When retrieving data, the stored ACTG sequences are decoded back into binary data, preserving the original digital information. This encoding and decoding mechanism facilitates seamless storage and retrieval of data in DNA form. Last layer of the multi-layer DNA storage system focuses on converting binary data into sequences of Adenine (A), Thymine (T), Cytosine (C), and Guanine (G), collectively known as ACTG sequences, which are essential for efficient encoding and retrieval of data in DNA molecules. This conversion process maps each binary digit (0 or 1) to its corresponding nucleotide base in DNA, bridging the digital and biological realms. During encoding, binary data is systematically translated into ACTG sequences to ensure accurate representation of each digit. When retrieving data, the stored ACTG sequences are decoded back into binary data, preserving the original digital information. This encoding and decoding mechanism facilitates seamless storage and retrieval of data in DNA form.

IV. Results and Discussion

This section aims to present the results of this research and implementation efforts, providing an overview of the significant outcomes of the approach presented in this study. This methodological journey has culminated in a pivotal moment, where the presentation of the results of the comparative study between four data compression techniques: Huffman Encoding, Run-Length Encoding (RLE), LZW (Lempel-Ziv-Welch), and LZ77 (Lempel-Ziv 1977), which form the foundation of the first layer of the proposed system. This critical comparison has shaped the core components of this approach. Additionally, a transition is made into the digital realm by introducing a user interface (UI), as shown in Fig. 2, which visually presents the holistic results of the study, providing a comprehensive view of the system's efficiency, data integrity, and ease of use.

The test aims to identify the most suitable compression technique for the first layer of this intelligent multi-layered solution. The proposed system also incorporates error correction and encryption, while considering the data's specifics and subsequent phases' requirements. Twitter data of various sizes was used, ranging from 1 MB to 50 MB, including positive and negative terms. About the amount of data, emphasis is placed on the need for small sizes for efficient storage and retrieval in DNA storage. Before the compression stage, this research compares various compression techniques to identify the one with the most optimal performance for this system. Thus, the same corpus of data was used with different algorithms. Fig. 3 shows an example of the files produced during compression and decompression by Huffman's algorithms, based on different file sizes. Other compression:

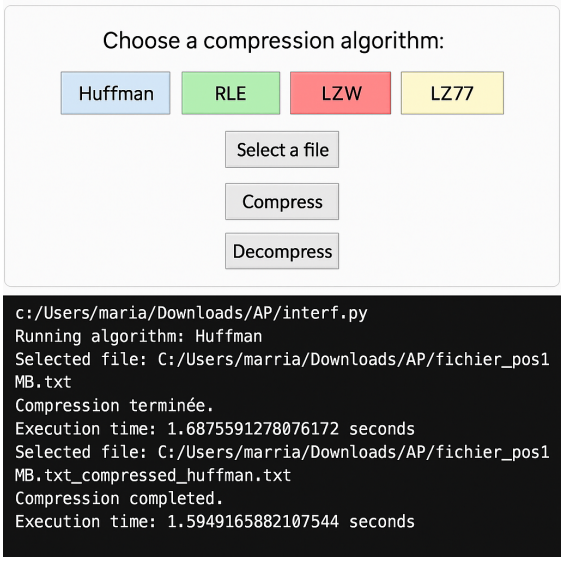

Fig. 2. Interface for running compression and decompression algorithms (e.g., Huffman 1MB result).

1) DNA Storage Process: .

Algorithms were also subjected to the same process. The following section presents an analysis and comparison of the results obtained. Examination of the results highlights the strengths and weaknesses of each compression technique while considering the particular needs of the developed multi-layer system. The performance and execution time associated with compression and decompression have been plotted below for each of the four compression algorithms (Huffman, RLE, LZW, LZ77), compiling and analyzing the results obtained.

10mb-explefile.txt_compressed_huffman	07/04/2024 12:56	Document texte	5 156 Ko
10mb-explefile.txt_decompressed_huffman	07/04/2024 13:01	Document texte	11 821 Ko
fi_50MB.txt_compressed_huffman	07/04/2024 11:25	Document texte	27 801 Ko
fichier_pos1MB.txt_compressed_huffman	07/04/2024 10:29	Document texte	48 883 Ko
fichier_pos1MB.txt_decompressed_huffman	07/04/2024 10:30	Document texte	585 Ko
fichier_pos5MB.txt_compressed_huffman	07/04/2024 10:31	Document texte	1 029 Ko
fichier_pos5MB.txt_decompressed_huffman	07/04/2024 10:33	Document texte	2 916 Ko
fichier_pos5MB.txt_decompressed_huffman	07/04/2024 10:33	Document texte	5 132 Ko

Fig. 3. Files generated by compression and decompression using the Huffman compression algorithm.

The results obtained when evaluating the four compression methods show that, in the context of DNA storage, the Huffman compression method produced the greatest size reduction for the different data sizes (1MB, 5MB, 10MB, 50MB), with a compression ratio varying between 40% and 50% (as shown in the Fig. 4). This result is particularly significant. This size gain is essential for DNA storage efficiency, where data size reduction is crucial before mapping onto ACGT sequences. Compared with other methods such as RLE and LZW, although RLE showed good results for highly redundant data, where compression and decompression time performance did not exceed 100 seconds, its performance declined with less regular data, such as that generated by tweets (Fig. 5). In Fig. 6, Fig. 7, LZW and LZ77, although less efficient than Huffman in terms of size reduction rates, stand out for their speed during

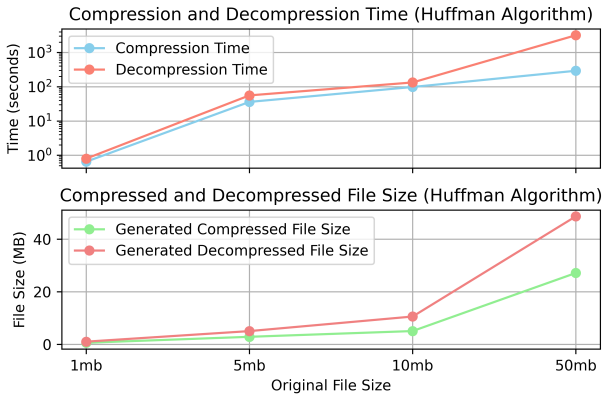

Fig. 4. Performance comparison of compression and decompression (Huffman algorithm).

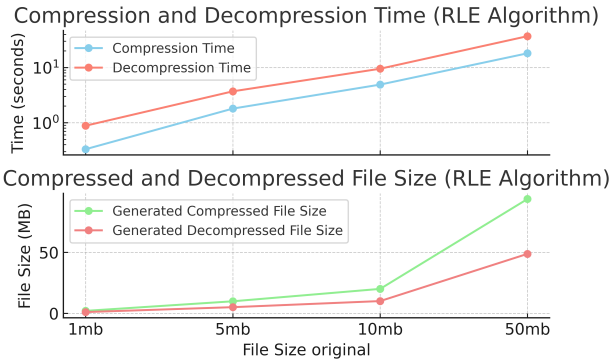

Fig. 5. Performance comparison of compression and decompression (RLE algorithm).

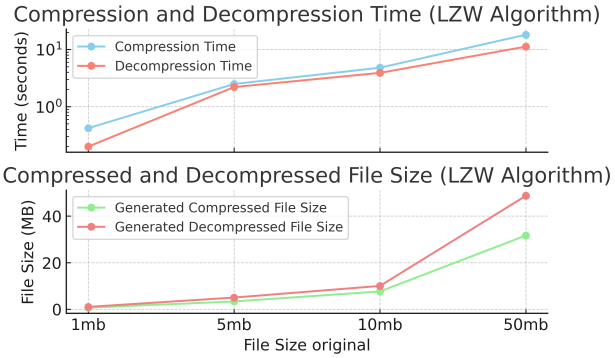

Fig. 6. Performance comparison of compression and decompression (LZW algorithm).

the compression and decompression processes. This makes them advantageous choices for contexts where processing time is a priority. This suggests that, although these methods are potentially suitable for other types of data, Huffman remains the most robust algorithm for this type of data, not least because of its ability to manage symbol frequency efficiently.

A comparative evaluation of the four compression techniques reveals significant disparities in terms of execution time and compression ratio. Although Huffman and LZW show

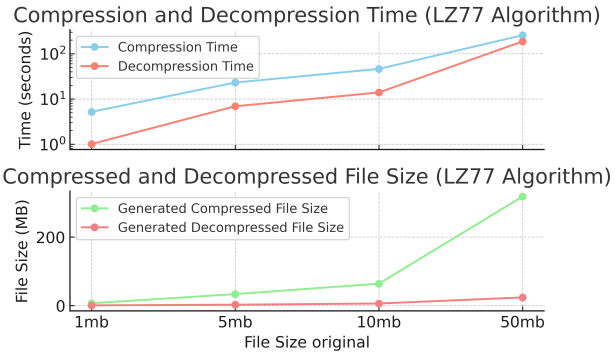

Fig. 7. Performance comparison of compression and decompression (LZ77 algorithm).

good compression ratios, their times differ, while LZ77 is less sophisticated in compression but stands out for its good performance during decompression. However, RLE is very fast and only adapts to data that is repeated. Thus, the choice of algorithm depends on the context: storage optimization (Huffman, LZW), fast transmission (LZ77), or management of repeated data (RLE).

One of the limitations of this study is that the performance of the compression methods was only evaluated on small to medium-sized data (up to 50MB). It would be relevant to conduct further tests on even larger datasets to verify whether Huffman's performance holds up with larger volumes. Furthermore, although the application of PGP encryption and polar coding in the pipeline ensured adequate security and data preparation for DNA synthesis, improvements in coding speed and error handling could be explored to further enhance the system's efficiency.

This research aims to overcome these obstacles by proposing a multi-level approach, focusing primarily on compression, a key factor in reducing costs and improving the efficiency of DNA storage. According to the results obtained, Huffman's algorithm offers a perfect balance between compression ratio, encoding and decoding speed, and preservation of data integrity. This implication helps to increase storage capacity and perfect the use of DNA storage resources. In summary, this study shows that the choice of compression algorithm is crucial for DNA storage, and this results underline the importance of an approach well adapted to the specific characteristics of the data. Future research should focus on optimizing the entire pipeline, integrating more advanced compression algorithms and testing their effectiveness on larger-scale datasets. This approach could open new avenues for faster, more reliable DNA storage.

V. Conclusion

In conclusion, this study explores the promising field of DNA storage and proposes a multi-layered approach to the challenges posed by conventional data storage methods. With data production reaching unprecedented levels, the demand for innovative and sustainable solutions is greater than ever. DNA storage is emerging as a compelling solution, offering remarkable density and durability. this system meets the need for a multi-layer solution combining DNA's unique storage

capabilities with advanced techniques such as data compression, error correction, and encryption. This work focuses on the first layer of this method, the compression layer, and through a comparative study of compression techniques, the effectiveness of Huffman coding for data compression is demonstrated. In addition, this study introduces a user interface for visualizing the results of our project, focusing on system efficiency, data integrity and user-friendliness. By combining biology and information technology, this approach paves the way for DNA storage to become a secure, reliable and efficient long-term solution for the preservation of large datasets. Optimization of compression algorithms and encryption methods could improve the efficiency of DNA storage. An advanced compression algorithm is used to minimize data size and optimize storage space. The aim is to create a robust and secure system for storing large quantities of data by developing a complete pipeline that combines compression, error correction and encryption. This method can also be applied to the storage of critical and sensitive data, particularly about biology, cybersecurity and the preservation of digital archives. Consequently, this research lays the foundation for a new intelligent DNA storage structure, capable of increasing storage efficiency and durability over the long term. In future studies, other layers will be analyzed to determine optimization methods and strategies for each component of the system.

ACKNOWLEDGMENT

This work is supported by National Center for Scientific and Technical Research (CNRST), Rabat, Morocco.

REFERENCES

[1] KL. Garner, "Principles of synthetic biology", Essays in Biochemistr, vol. 65, no. 5, pp. 791–811, 2021.

[2] PY. Silva, GU. Ganegoda, "New trends of digital data storage in DNA", BioMed Research International, 2016.

[3] O. Jelezniak, "Digitalization, digitization of the world and digital detox", Project Baikal, vol. 71, pp. 92–99, 2022.

[4] L. Ceze, J. Nivala, K. Strauss, "Molecular digital data storage using DNA", Nat Rev Genet, vol. 20, pp. 456–466, 2019.

[5] G. Church, Y. Gao, S. Kosuri, "Next-Generation digital information storage in DNA", Science, vol. 337, 2012.

[6] Y. Erlich, D. Zielinski, "DNA fountain enables a robust and efficient storage architecture", Science, vol. 355, no. 6328, pp. 950–954, 2017.

[7] Y. Nahum, E. Ben-Tolila, L. Anavy, "Single-Read reconstruction for DNA data storage using transformers", arXiv: Emerging Technologies, 2021.

[8] S. Karthikeyan, T. Poongodi, "Secure data transmission in smart cities using DNA cryptography with LZW compression algorithm", Optoelectronics, Instrumentation and Data Processing, vol. 60, pp. 156–167, 2024.

[9] Ch. Jiang, Y. Zhang, F. Wang, H. Liu, "Toward smart information processing with synthetic DNA molecules", Macromolecular Rapid Communications, vol. 42, no. 11, 2021.

[10] P. Hofmann, JA. Cabrera, E. Krieg, R. Bassoli, F. Fitzek, "DNA-Storage in Future Communication Networks", IEEE Communications Magazine, pp. 1–7, 2023.

[11] K. Z. Abram, Z. Udaondo, "Leveraging nature to advance data storage: DNA as a storage medium", Microbial biotechnology, 2023.

[12] I. Shomorony, R. Heckel, "DNA-Based storage: models and fundamental limits", IEEE Transactions on Information Theory, vol. 67, no. 6, pp. 3675–3689, 2021.

[13] Y. Zheng, B. Cao, J. Wu, B. Wang, Q. Zhang, "High net information density DNA data storage by the mope encoding algorithm", IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1–10, 2023.

[14] T. Xu, P. Zhou, "Feature extraction for payload classification: a byte pair encoding algorithm", IEEE 8th International Conference on Computer and Communications, Dec. 2023.

[15] D. Bar-Lev, "Deep DNA storage: scalable and robust DNA storage via coding theory and deep learning", arXiv: Information Theory, 2021.

[16] S. Samir, M. Ramdane, M. Lalam, M. Ahmed-Ouamer, "Algorithm for data compression", Electronic Journal of Information Technology, 2007.

[17] A. Moffat, "Huffman coding", ACM Computing Surveys, pp. 1–35, 2019.

[18] S. Fiergolla, "Improving run length encoding by preprocessing", arXiv, Data Structures and Algorithms, 2021.

[19] Y. He, X. Shi, Y. Wang, "A Modified LZW algorithm based on a character string parallel search in cluster-based telemetry data compression", Electronics, vol. 11, pp. 2656–2656, 2022.

[20] A. Hong, M. Rossi, C. H. Boucher, "LZ77 via prefix-free parsing", Workshop on Algorithm Engineering and Experimentation, pp. 123–134, 2023.

[21] H. B. Abdalla, "A brief survey on big data: technologies, terminologies and data-intensive applications", Journal of Big Data, vol. 9, no. 107, 2022.

[22] R. Beakta, "Big Data and Hadoop: a review paper", International Journal of Computer Science & Information, vol. 2, 2015.

[23] U. R. Raje, "A comparative study of data storage in retail management with traditional databases v/s real time database", International Journal of Advanced Research in Science, Communication and Technology, pp. 307–310, 2022.

[24] S. Weber, "Data, development, and growth", Business and Politics, vol. 9, no. 3, pp. 397–423, 2017.

[25] S. Kaisler, F. Armour, J. A. Espinosa, W. Money, "Big Data: issues and challenges moving forward", 46th Hawaii International Conference on System Sciences, Jan. 2013.

[26] A. Hong, W. Xiao, J. Ge, "Big Data analysis system based on Cloudera distribution Hadoop", Int. Conf. on Big Data Security on Cloud, IEEE, pp. 169–173, May 2021.

[27] M. Sarwat, S. Elnikety, Y. He, M. F. Mokbel, "Horton+: a distributed system for processing declarative reachability queries over partitioned graphs", Proc. VLDB Endowment, vol. 6, no. 14, pp. 1918–1929, 2013.

[28] A. Modi, R. Sha, K. Jain, R. Verma, R. Shorey, H. Saran, "Multi-Agent packet routing (MAPR): co-operative packet routing algorithm with multi-agent reinforcement learning", 15th Int. Conf. on Communication Systems & Networks, Jan. 2023.

[29] E. A. Jrai et al., "Improving LZW compression of Unicode Arabic text using multi-level encoding and a variable-length phrase code", IEEE Access, vol. 11, pp. 51915–51929, 2023.

[30] R. Xie et al., "Study of the error correction capability of multiple sequence alignment algorithm (MAFFT) in DNA storage", BMC Bioinformatics, vol. 24, no. 111, 2023.

[31] B. J. Cafferty et al., "Storage of Information Using Small Organic Molecules", ACS Central Science, vol. 5, no. 5, pp. 911–916, 2019.

[32] A. A. Nagarkar et al., "Storing and reading information in mixtures of fluorescent molecules", ACS Central Science, vol. 7, no. 10, pp. 1728–1735, 2021.

[33] L. Yu et al., "Digital synthetic polymers for information storage", Chemical Society Reviews, vol. 52, pp. 1529–1548, 2023.

[34] L. Organick et al., "Probing the physical limits of reliable DNA data retrieval", Nature Communications, vol. 11, 2020.

[35] Z. Ping et al., "Carbon-based archiving: current progress and future prospects of DNA-based data storage", GigaScience, vol. 8, 2019.

[36] D. Sharma, M. Ramteke, "Chapter Seven - DNA computing-based Big Data storage", Advances in Computers, vol. 129, pp. 249–279, 2023.

[37] M. E. Allentoft et al., "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils", Proc. Royal Society B: Biological Sciences, vol. 279, pp. 4724–4733, 2012.

[38] E. Bencurova et al., "DNA storage—from natural biology to synthetic biology", Computational and Structural Biotechnology Journal, vol. 21, pp. 1227–1235, 2023.

[39] J. D. Watson, F. H. Crick, "The structure of DNA", Cold Spring Harbor Symposia on Quantitative Biology, vol. 18, pp. 123–131, 1953.

[40] D. Lavenier, "DNA Storage: Synthesis and Sequencing Semiconductor Technologies", Int. Electron Devices Meeting, Dec. 2022.

[41] S. Ray, "DNA data storage", Medium, 2021.

[42] S. Alkhliwi, "Huffman encoding with white tailed eagle algorithm-based image steganography technique", Eng., Tech. & Appl. Sci. Research, vol. 13, no. 2, pp. 10453–10459, 2023.