Biological constraint in digital data encoding: A DNA based approach for image representation

Muhammad Rafi Muttaqin [a,b,1], Yeni Herdiyeni [a,2,*], Agus Buono [a,3], Karlisa Priandana [a,4], Iskandar Zulkarnaen Siregar [c,5], Wisnu Ananta Kusuma [a,6]

[a] Department of Computer Science, IPB University, Bogor, Indonesia
[b] Informatic Engineering, Sekolah Tinggi Teknologi Wastukancana, Purwakarta, Indonesia
[c] Department of Silviculture, IPB University, Bogor, Indonesia
[1] rafiaqinmuttaqin@apps.ipb.ac.id; [2] yeni.herdiyeni@apps.ipb.ac.id; [3] agusbuono@apps.ipb.ac.id; [4] karlisa@apps.ipb.ac.id;
[5] siregar@apps.ipb.ac.id; [6] ananta@apps.ipb.ac.id
* corresponding author

ARTICLE INFO

Article history
Received September 10, 2024
Revised June 24, 2025
Accepted June 27, 2025
Available online July 26, 2025

Keywords
Biological constraint
Digital image encoding
DNA data storage
MNIST dataset
Multiple sequence alignment

ABSTRACT

Digital data encoding is crucial for communication and data storage, but conventional techniques, such as ASCII and binary coding, have drawbacks in terms of processing speed and storage capacity. A potential substitute with parallel processing and high-capacity storage is DNA-based data encoding. The goal of this research is to develop a digital data encoding technique based on DNA, while considering biological constraints such as homopolymer and GC-content. The process involves converting image pixel values into binary format, followed by encoding into DNA sequences, ensuring they meet biological constraints. The validity of the resulting DNA sequences is assessed through transcription and translation processes. Additionally, Multiple Sequence Alignment analysis is conducted to compare the similarities between the encoded DNA sequences. The results indicate that the DNA sequences from MNIST images share similar characteristics, reflected in the phylogenetic tree's close clustering. Multiple Sequence Alignment analysis shows that biological constraints successfully preserved the core visual features, allowing accurate clustering. However, this method also faces drawbacks, particularly in the reduction of visual information and sensitivity to changes in image intensity. Despite these challenges, DNA-based encoding shows potential for digital image representation. Further development, particularly the integration of deep learning, could lead to more efficient, secure, and sustainable data storage systems, especially for image data.

1. Introduction

Digital data encoding is a key concept in our technologies for data storage, information, and communication networks. It includes transforming data that can be used for transmission, storage, or interpretation using computer systems. Binary and ASCII encoding have been widely used as conventional encoding methods. Unfortunately, these methods limit the storage capacity and speed of data processing [1]–[3]. Along with the rapid development of digital data usage, there is a considerable need for more effective data encoding techniques. Due to their effectiveness, traditional image encoding techniques such as PNG and JPEG have been widely employed. However, this approach has limitations because some multimedia applications cannot process them properly, thus requiring other approaches. Alternative encoding techniques, including DNA-based data storage, have been researched by several

researchers [4], [5]. Several researchers have researched alternative encoding techniques, including DNA-based data storage [6].

DNA encoding is the process of converting digital data into a sequence of the basic building blocks of DNA, namely adenine (A), thymine (T), cytosine (C), and guanine (G). The main advantage of this approach for long-term data storage is its density and durability [7], [8]. If DNA is stored correctly, it can remain stable for thousands of years, allowing it to be used to produce images. DNA parallel processing capabilities can lead to significant progress in data analysis and computational tasks [9]. Thus, this method for storing data and pictures is a renewable technology [10]. Digital data is mapped to DNA sequences to encode DNA-based data [11], [12]. This is done by dividing digital data into smaller segments, synthesizing the corresponding DNA strands, and assigning specific DNA sequences to represent the data [13], [14]. Several encoding techniques have been developed to effectively encode digital information into DNA molecules, such as utilizing specific nucleotide sequences to represent binary data. Following decoding, the obtained DNA sequence is mapped back to the original digital data [15]. Sequencing the DNA strand and restoring the nucleotide sequence to its original digital format are steps in decoding. This procedure often employs high-throughput sequencing technology and computational techniques to recover encoded data from DNA molecules precisely [16], [17].

DNA-based data encoding presents several challenges, including developing effective data retrieval techniques, ensuring error-free encoding and decoding, and the precise and efficient synthesis and sequencing of DNA [18], [19]. However, overcoming these challenges will open up opportunities for improvements in computing, communication, and data storage. The biological constraints of this procedure can affect the accuracy of the sequence data obtained. The DNA synthesis, sequencing, and decoding processes face two significant obstacles: homopolymer repeats and GC content bias. These factors ultimately affect the accuracy of genetic data obtained from sequencing [13], [20], [21]. Repeating a single base or a series of identical bases within a DNA molecule results in homopolymeric repeats. During DNA sequencing, repeat sequences such as “AAAA” or “TTTT” make it difficult for DNA polymerase enzymes to determine the exact number of repeats [12]. Consequently, sequence analysis may be inaccurate if a sequence containing four adenine (A) repeats is read as three or five repeats instead of four. DNA bond formation may also be affected by homopolymeric sequences, especially during PCR amplification. Errors in DNA amplification may arise due to DNA polymerase enzyme slippage caused by long homopolymeric sequences.

The GC content tendency is defined as an imbalance in the distribution of guanine (G) and cytosine (C) base pairs compared to adenine (A) and thymine (T) [22]. DNA sequence results can be influenced by the instability of GC and AT base pairs. This is because guanine and cytosine form stronger hydrogen bonds than adenine and thymine. GC concentration is crucial for DNA synthesis, making DNA sequences with higher GC content more stable. The limitations of DNA are significant in fields such as molecular biology, genetics, and biotechnology [22]. Moreover, the manipulation, analysis, and understanding of DNA are significantly influenced by limitations such as GC content bias and repetitive nucleotide sequences.

Errors or inconsistencies may arise during DNA sequencing, synthesis, and data interpretation due to biological constraints [23]. Resolving these issues is crucial for maintaining accurate and reliable DNA data, which is essential for gene editing, diagnosis, and personalized medical interventions. DNA data can be misinterpreted due to an inability to properly account for biological constraints. In particular, incorrect genome assembly or genetic variant identification may result from misreading homopolymeric regions or failing to account for GC content bias. Considering these limitations, researchers can ensure the validity of their findings and avoid misunderstandings. In addition, Researchers may enhance their experimental protocols in DNA manipulation and analysis by understanding the biological constraints and how to overcome them. For instance, knowledge of homopolymer regions facilitates the development of primers for PCR amplification, and awareness of GC content bias can inform the choice of platforms and sequencing processes.

By overcoming biological constraints, researchers may enhance the overall quality of experimental DNA data. It could be accomplished by decreasing background noise in genetic analysis, minimizing sequencing errors, and increasing the signal-to-noise ratio of DNA sequencing data. Furthermore, high-quality DNA data is necessary in biological research to draw insightful conclusions and create accurate insights. Efforts to overcome biological constraints drive breakthroughs in DNA analysis techniques and technologies. Researchers invented advanced sequencing platforms, algorithms, and bioinformatics tools to overcome issues such as homopolymer repetition and GC bias. These developments are driving the advancement of genomics and biotechnology, helping to improve current procedures. Suppose researchers can use reliable methods to mitigate biological constraints, such as homopolymer regions and GC bias, while improving the reproducibility standard of analysis. In that case, the results of DNA experiments from various laboratories will be more consistent. Overall, overcoming biological constraints of DNA determines the accuracy, reliability, and reproducibility of genetic data [22], [24]. Understanding these constraints and devising strategies to overcome them enables researchers to develop innovative genetic research, refine diagnostic and therapeutic applications, and stimulate innovation in molecular biology and biotechnology.

DNA-based digital data encoding is currently gaining traction. Researchers across disciplines constantly experiment with various methods of utilizing DNA as a storage medium for digital information. Their goal is to develop more efficient and cost-effective data encoding and reading techniques. These efforts involve exploring alternative encoding algorithms, error repair mechanisms, and DNA synthesis technology. Their goal is to ensure that data is accurately and reliably stored in DNA molecules. In addition, previous studies have extensively researched the potential applications of DNA-based data storage, such as storage media, data security, and information preservation [25]–[27]. Dorrichi [28] stated that DNA was chosen because of its ability to store a large amount of information stably. This promising idea addresses the growing challenges of digital data storage. The long-term stability of DNA molecules makes them an ideal choice for protecting important data for future generations [29]. Regulations and ethics related to DNA encoding digital data have also been evaluated. This is because using biological materials for data storage is closely related to information access, privacy, and security.

The density and natural durability of DNA molecules make DNA-based data storage attractive. This capability can provide high security against potential cyber threats and data degradation. The result will be secure and impenetrable data storage. DNA-based digital data encryption systems could completely change how digital information is stored and preserved. DNA-based data storage addresses the weaknesses of conventional storage devices such as magnetic tape and hard drives. Preserving historical and cultural documents is an additional advantage in information preservation. The considerable storage capacity of DNA allows for long-term and sustainable storage of information archives. Previous researchers have studied DNA encoding, such as the studies by [30] and [31], which examined storage capacity efficiency and the design of optimal DNA encoding algorithms. There are also studies by [22], [32] and [20] that investigated GC content control and homopolymer control, but their applications were limited to text or binary data, not images. However, these studies still only focused on mapping digital data to DNA sequences by optimizing encoding density or error correction efficiency. Therefore, this study focuses on an approach that explicitly considers biological constraints at each encoding stage, not just the technical aspects. Additionally, it incorporates transcription and translation processes to validate biocompatibility, which is rarely done in previous DNA encoding studies. This study also uses image data as the test domain, not just text or binary data, but involves the challenge of representing more complex information. The multiple sequence alignment (MSA) process is integrated to assess the compatibility between the generated sequences and evaluate how well the visual features of the image are preserved in the DNA representation. Thus, this study opens up the possibility of end-to-end DNA encoding for image-based data, which has been relatively unexplored in the literature. It is also hoped that the general public understand the added value of this approach.

DNA-based coding in this work focuses mostly on the representation of digital images in general. The MNIST handwriting dataset, a set of basic digital pictures of handwritten numbers, was utilized as

a preliminary step and for testing. The MNIST dataset was selected to learn how the visual shape of an image influences DNA sequence encoding outcomes. The MNIST dataset allows us to investigate how image shape variability affects the final coding outcomes, including the stability of the generated sequences and the suitability of phylogenetic clustering. The impact of image shape variability on the final coding outcomes, including the adequacy of phylogenetic clustering and the stability of the generated sequences, can be investigated using the MNIST dataset.

The findings of this study guarantee that this strategy could potentially be modified and broadly used to many kinds of visual data, building a strong basis for future applications of the same coding method to increasingly complex digital images.

2. Method

This research has some steps, as shown in Fig. 1. The process begins by entering the original image, which then undergoes binarization to convert it into binary data. This binary data is then encoded into a DNA sequence (Encoding to DNA Based [A, T, C, G]) while considering biological constraints to ensure the encoding is valid. Following this, validation is performed. If the result is invalid, the process returns to the DNA encoding stage for correction. If valid, the DNA data is processed to the evaluation stage, then continued with clustering using multiple sequence alignment for further analysis.

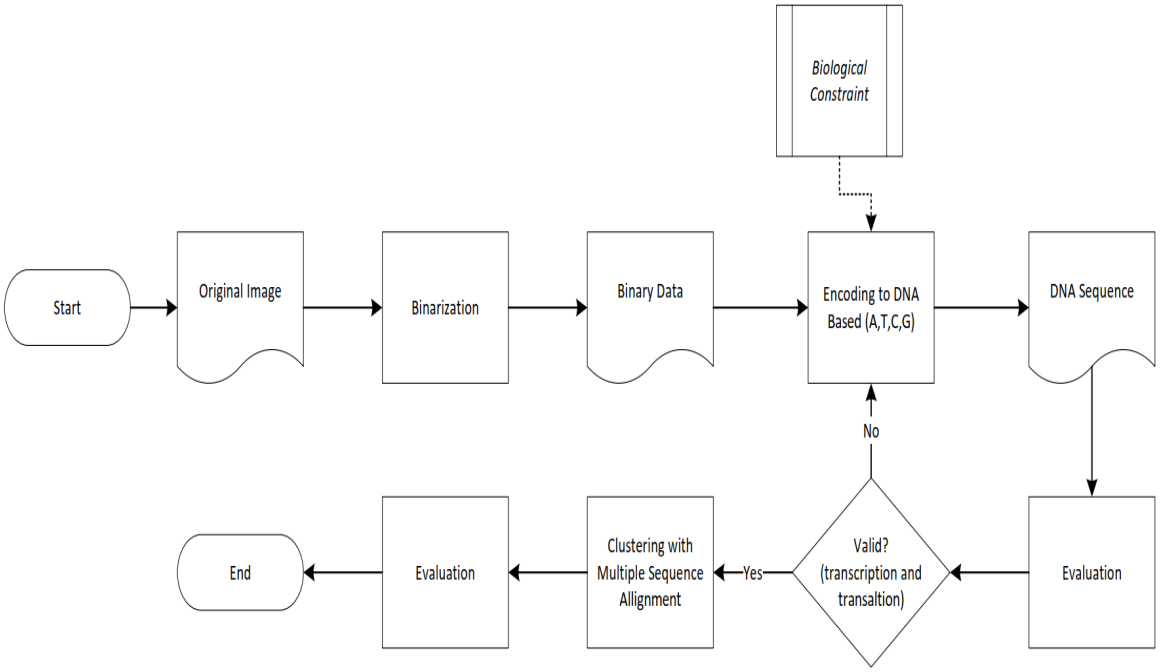

Fig. 1. Research Methodology

2.1. Binarization

The image, containing pixel intensity values in decimal form, was converted into a binary format with an 8-bit length for each pixel intensity value. Therefore, if the intensity value of the image was 255, it was converted into a binary format of 11111111. For example, a pixel intensity value of 125 was converted to 01111101. This decimal-to-binary encoding process outputs image data containing values in a binary format. This binary format is processed into a DNA sequence, considering biological constraints during the encoding stage.

2.2. DNA Encoding

The steps in the algorithm used for binary-to-nucleotide mapping and homopolymer constraints are outlined as in Algorithm 1.

Algorithm 1

In this process, each pixel of the grayscale image is converted into an 8-bit binary.

Step 1: The binary sequence is processed 2 bits per step to generate a nucleotide code based on the following rules:

00 A
01 C
10 G
11 T

Step 2: To prevent the formation of homopolymers beyond a threshold, when a run-length of 3 identical bases is detected, the subsequent encoding process will force base switching with conditional rules:

a. If the last base is A or T, the next bit is encoded as C or G
b. If the last base is C or G, the next bit is encoded as A and T

Step 3: Read the next two bits and repeat Step 1

The encoding algorithm with homopolymer constraints is as follows [33], for example, the threshold for homopolymer formation is three consecutive base pairs.

```
Binary data: 0101010101010
Step 1: DNA: C C C
Step 2: The next bit is 0, then the next base is A (0:A)
        DNA: C C C A
Step 3: DNA: C C C A G G G
```

2.3. GC-Content Adjustment

The algorithm for handling GC-content that exceeds the predetermined limit or threshold is as in Algorithm 2.

Algorithm 2

i. Calculate the proportion of the GC base with the formula: GC-Content = $\frac{\sum G+\sum C}{\sum(G,C,A,T)}$

ii. If it meets the threshold, then no other bases should be added.

If it does not meet the threshold, then add the "appropriate" base so that the GC proportion reaches the desired limit.

For example, the specified GC content threshold is 40–60% [34].

DNA Data : C C C A G G G
GC-content : 85%
Calculate : $0.4 \leq \frac{6}{(7+x)} \leq 0.6$
: $3 \leq x \leq 8$
Add base : C C C A G G G A T A
Latest GC-content: 60%

In this coding algorithm, homopolymer constraints are applied to nucleotides during the 2-bit mapping process by monitoring the length of identical base sequences. If a sequence length of three or more is detected, the following bit mapping is forced to switch to a different base using predefined rules (A/T → C/G, C/G → A/T). It is to prevent homopolymer sequences from exceeding three nucleotides. Meanwhile, GC content constraints are applied after the complete sequence is generated by calculating the GC ratio. If the ratio is outside the 40–60% range, G or C bases are added to bring the ratio within the target range. It is to ensure that the sequence structure remains intact.

The output of the encoding process for the DNA sequence is DNA sequence data free from homopolymers and GC-Content. These DNA sequence data will be evaluated first to determine whether they can be translated into a protein. This evaluation is required to ensure that DNA sequences derived from digital images or data can be biologically bound if the synthesis process is performed

2.4. Evaluation: Transcription and Translation

DNA-based storage ensures biological systems can synthesize and process biologically generated DNA sequences. Therefore, in the evaluation phase, the resulting DNA sequence will be validated to determine

whether it is by the provisions that can form DNA bonds if the synthesis process is carried out. DNA sequence transcription and translation were used as validation methods. Transcription and translation act as biological sanity checks to ensure the resulting DNA sequence does not contain toxic elements or unstable sequences. Transcription and translation ensure that the resulting DNA sequence can also form valid mRNAs and proteins, making it potential for applications in bio-storage that require integration with biological or hybrid systems. Thus, this process is not simply an additional step but an integral part of verifying the biocompatibility of the encoded DNA sequence.

The main principle of gene expression consists of two consecutive steps: transcription (DNA to mRNA) and translation (mRNA to protein). The transcription and translation processes are shown in Fig. 2. Transcription is the main step that controls whether a gene is active and determines the identity and status of the cell.

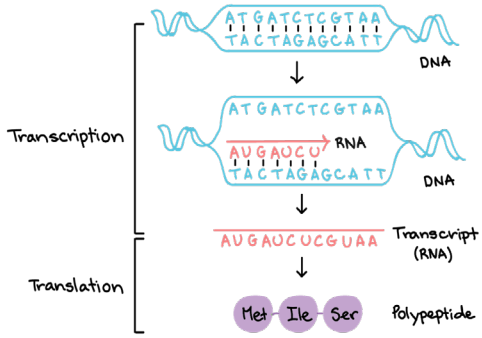

Fig. 2. Transcription and Translation Process

Protein formation, also known as protein synthesis, involves two main steps: transcription and translation (Fig. 3). The application of transcription and translation in this study evaluates the safety and biological feasibility of the DNA produced, not merely as a symbolic representation of genetic structure. This process is biologically motivated, not intended to mimic biological structures. The aim is to ensure that the DNA sequence resulting from encoding can be validly transcribed and translated into protein, as in real biological systems. This validation is crucial because, in the context of long-term DNA-based data storage, the stability and biocompatibility of the sequence are essential to ensure the feasibility of synthesis and maintain data integrity.

Transcription is the first step in protein synthesis. During transcription, DNA sequences are converted into complementary RNA sequences. This process occurs in the nucleus of eukaryotic organisms and the cytoplasm of prokaryotic organisms. Transcription is carried out by an enzyme called RNA polymerase, which binds to a specific region of DNA called the promoter. RNA polymerase unwinds the DNA double helix and uses one strand of DNA as a template to synthesize a complementary RNA strand. This complementary RNA, known as messenger RNA (mRNA), carries genetic information from the DNA to the ribosome, where protein synthesis occurs. The complementary rules from DNA to RNA are listed in Table 1.

Table 1. Complementary DNA to RNA

DNA	RNA
A	U
T	A
G	C
C	G

Once the transcription process is complete, the translation process continues. The translation process reads the mRNA to produce proteins. The mRNA molecule is a single strand, and its base sequence consists of adenine (A), uracil (U), cytosine (C), and guanine (G), which are complemented by DNA bases. Each mRNA molecule specifies the amino acid sequence that must be added for protein formation.

Three nucleotides or codons represent each amino acid in the mRNA molecule. For example, AGC is the mRNA codon for the amino acid serine, and UAA is the signal that stops translation. The rules of protein representation based on the nucleotides of an mRNA molecule are shown in Fig. 3.

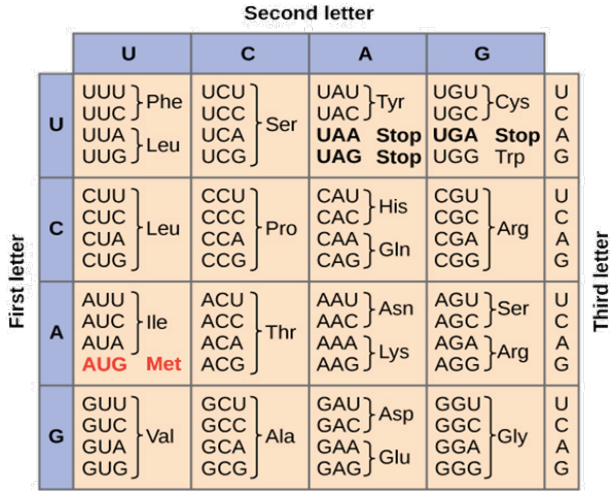

Fig. 3. Rules of mRNA to protein conversion (https://openstax.org/books/biology/pages/15-1-the-genetic-code)

The amino acids from this translation process are Phenylalanine (Phe), Leucine (Leu), Serine (Ser), Tyrosine (Tyr), Cysteine (Cys), Tryptophan (Trp), Pronline (Pro), Histidine (His), Glutamine (Gln), Arginine (Arg), Lysine (Lys), Threonine (Thr), Isoleucine (Ile), Methionine (Met), Valine (Val), Alanine (Ala), Glutamate (Glu), Aspartate (Asp), Glysine (Gly), and Asparagine (Asn) [35] (Fig. 4).

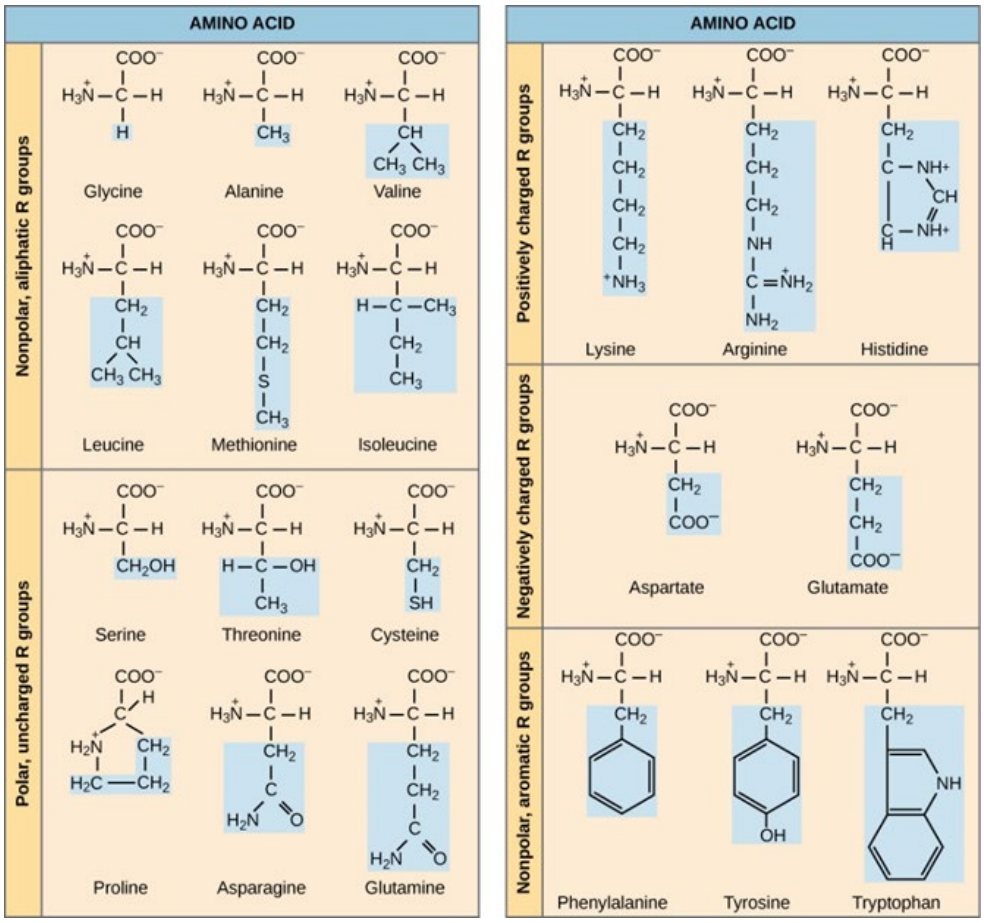

Fig. 4. Proteins produced from mRNA (https://openstax.org/books/biology/pages/15-1-the-genetic-code)

The illustrative example of the process of converting amino acids/three mRNA bases into proteins is shown in Fig. 5. If all the DNA sequences can be converted into proteins, it can be assumed that the DNA sequence data can form biological bonds.

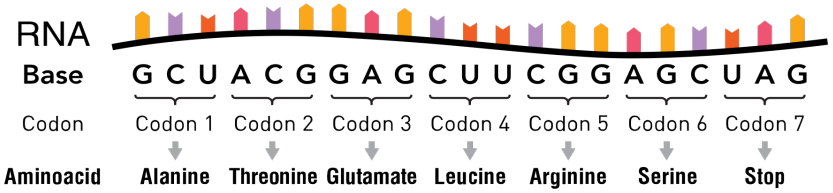

Fig. 5. Illustration of formation from RNA to Protein Amino Acids (https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/translation/a/translation-overview)

The DNA sequence that has been evaluated is then subjected to a clustering process using the multiple sequence alignment method to determine the level of similarity of DNA sequences derived from several classes of datasets [36]. The purpose of this MSA is to determine whether there are similarities in the DNA sequence results from several dataset images that come from the same class. The results of this clustering will be evaluated, and some conclusions will be drawn.

3. Results and Discussion

3.1. Dataset used

This study used the MNIST Digital Handwriting dataset. The MNIST dataset contains 70,000 (seventy thousand) handwritten digit images. Each image in MNIST has a size of 28 × 28 pixels and is in a grayscale format. The pixel values varied from 0 (black) to 255 (white). MNIST was used in this study for several reasons. Due to its simplicity and ease of use, MNIST offers a robust baseline for machine learning experiments, enabling researchers to concentrate on algorithm development without concerns about data quality. Therefore, researchers should focus on testing and validating these algorithms. In this study, five pieces of data were used for each class, totalling 50 images.

3.2. Encoding MNIST Data into DNA Sequence

Encoding an MNIST image into a DNA sequence begins with the binarization process, which converts the pixel values in an image from a decimal format to a binary format. One MNIST image has 784-pixel values in decimal format. The format is converted into binary values; thus, it will have 6272 binary values. Each decimal value was converted into eight binary digits. For example, a value of 255 pixels was converted into eight binary digits: 11111111. A value of 100 pixels was converted to 01100100. After the image data is converted to binary data, encoding the DNA Sequence begins with the algorithm presented in section two. The stages for converting an image into a DNA sequence are shown in Fig. 6.

In Fig. 6, the image of the handwritten MNIST digit' 0' in grayscale format has an intensity value between 0 and 255 in decimal format. Where 0 represents black and 255 represents white. The number of digits in the image intensity values is 784. The decimal value of the intensity was converted into an eight-digit binary value for every decimal value. Thus, 784 decimal values were converted into binary values of 6272 digits. A set of binary numbers is encoded into a DNA Sequence using an encoding algorithm [37]. The encoding algorithm overcomes the limitations of homopolymers. Three base repetitions were performed in this study, which is three pieces. Thus, the results obtained were free from homopolymer restrictions. From the results obtained, it can be seen that the GC content limit was not met, as it fell below the specified threshold of 40%–60%. The GC value of the resulting DNA sequence was 32.58%, attributed to the number of G bases (354) and C bases (668) among a total of 3136 bases. Therefore, for the GC value to be within the predetermined threshold range, the G and C bases are added to the DNA Sequence generated from the encoding process. After adding the G and C bases, the GC-Content value was 40%, with 548 G bases and 862 C bases out of 3524 bases.

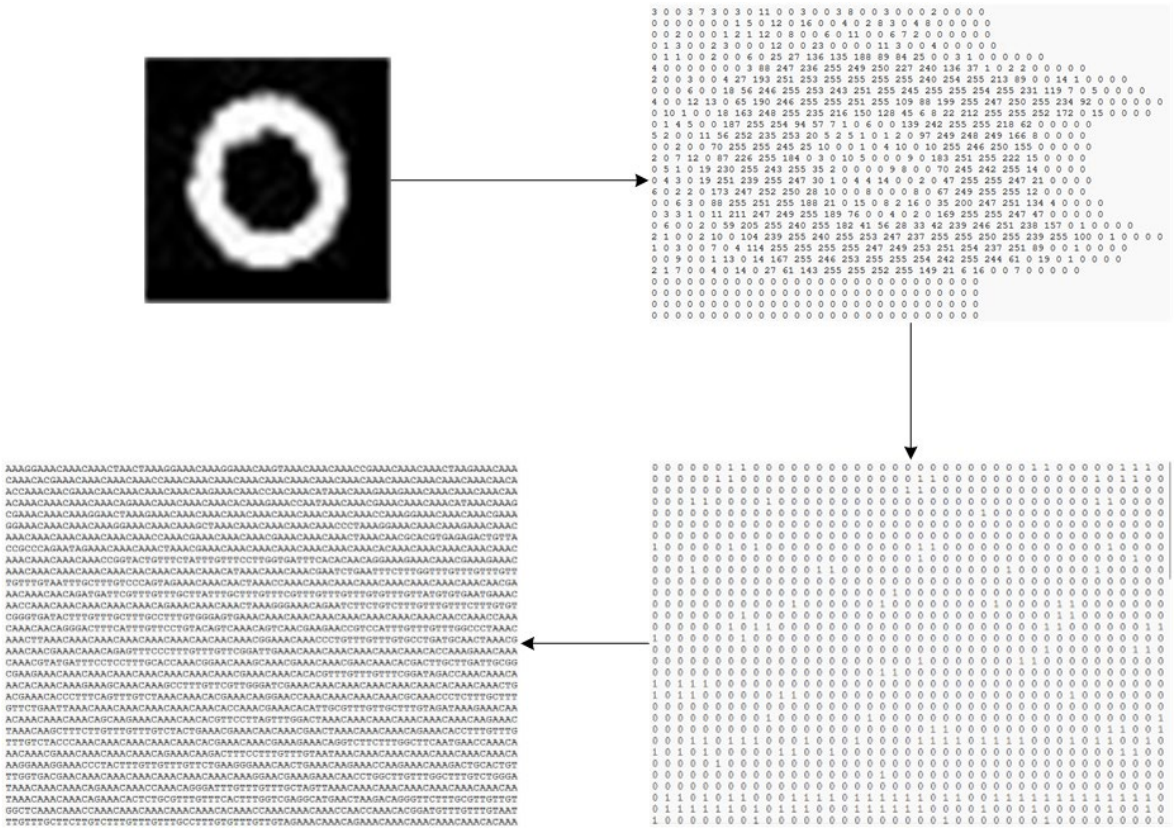

Fig. 6. The process of a MNIST image being converted into a DNA Sequence

3.3. Transcription of DNA Sequence into mRNA

The resulting DNA sequence is converted into mRNA, which is then converted back into amino acid proteins via a process called transcription. This process changes each base in the DNA Sequence according to its complementary base pairs. The results of the transcription process are shown in Fig. 7. We have attempted to categorize them into three bases to facilitate understanding of the subsequent process: translation.

Fig. 7. Transcription process from the DNA sequence image 0 MNIST into mRNA form

This transcription process produces mRNA that differs from its DNA sequence. Base A is converted into U, T with A, C with G, and G with C. The number of bases in the mRNA sequence was the same as that in the DNA Sequence.

3.4. Translation of mRNA into Protein Amino Acids

After mRNA is formed, it is translated into proteins made up of amino acids. Amino acid proteins are produced from three mRNA bases, commonly called codons. These codons form amino acid proteins

according to biological rules (Fig. 3). The process of converting DNA sequences into amino acid proteins is crucial because it ensures that genetic information is expressed correctly and functionally. These findings have broad implications for human health, scientific research, and biotechnology. The results of mRNA translation into amino acid proteins are shown in Fig. 8.

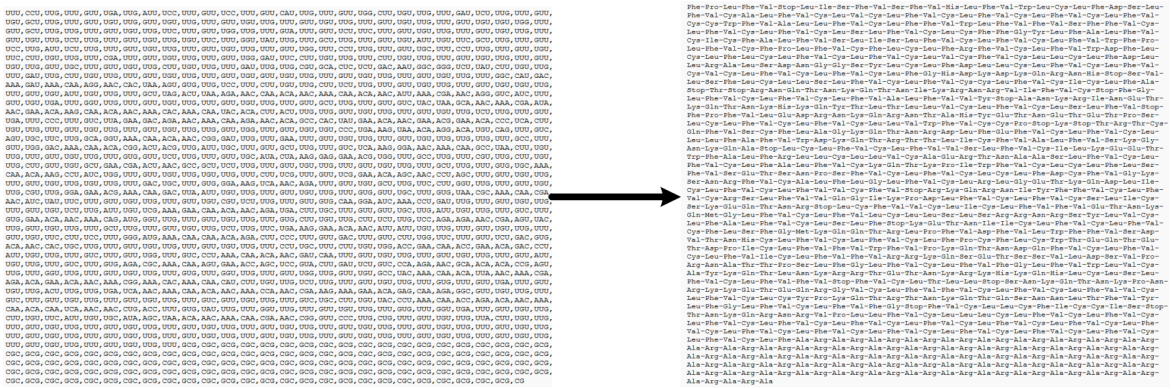

Fig. 8. Results of the mRNA-to-protein translation of amino acids

3.5. Multiple Sequence Alignment

We experimented with the form of multiple sequence alignment (MSA) of DNA sequences derived from the MNIST image encoding results [19], [38]. Using this method, we observed similarities and conservation patterns between the sequences. Five images were obtained for each class/number. The classes we took were numbered “zero”, “one”, “two”, “three”, and “four”. Fig. 9 shows the image that is the object of the MSA experiment resulting from encoding the DNA sequence. The label “Image 0-1” is the image of class zero sequence number one.

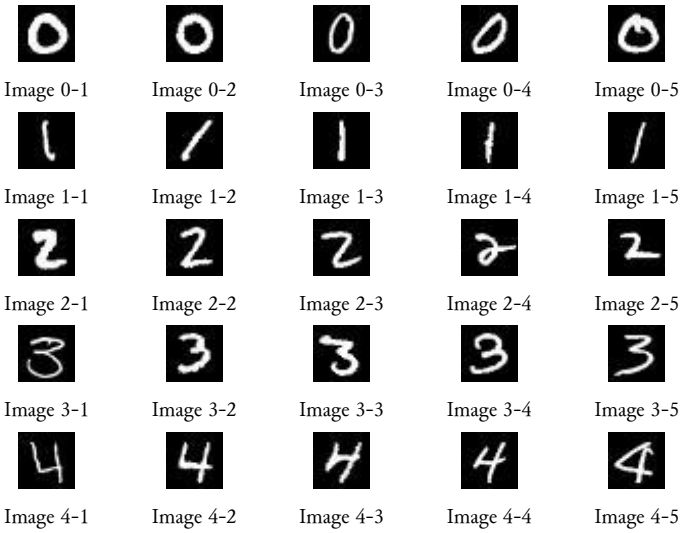

Fig. 9. Datasets used in multiple sequence alignment analysis.

This study utilizes MSA and phylogenetic trees to evaluate similarity based on DNA sequences, with the sequence similarity score employed as the primary quantitative indicator. Encoding images into DNA produces sequential representations that can reflect the continuity of key visual features. MSA can identify and visualize conserved regions between sequences, indicating similar visual patterns. Meanwhile, phylogenetic trees map distances between sequences into meaningful visual distances. Visual features are successfully preserved in the DNA domain when sequences from the same image class are clustered together. MSA compatibility score and phylogenetic tree structure represent the degree of preservation of visual features between encoded DNA sequences.

The MSA analysis conducted in this study used the CLUSTALW application (Kyoto University Bioinformatics Centre). The parameters used in MSA in the CLUSTALW application are a gap open penalty of 15 and a gap extension penalty of 6.66. The Weight Matrix used is the International Union of Biochemistry (IUB) method. The results are shown in Fig. 10. Fig. 10 displays a phylogenetic tree showing the similarity between the DNA sequences generated from the MNIST image encoding classes 0–4. Qualitatively, this tree illustrates how sequences from the same class are grouped closely, indicating strong similarities. The short branches in this tree indicate little difference between the sequences grouped. This indicates that the visual features in images of the same class are highly similar after being encoded into DNA sequences.

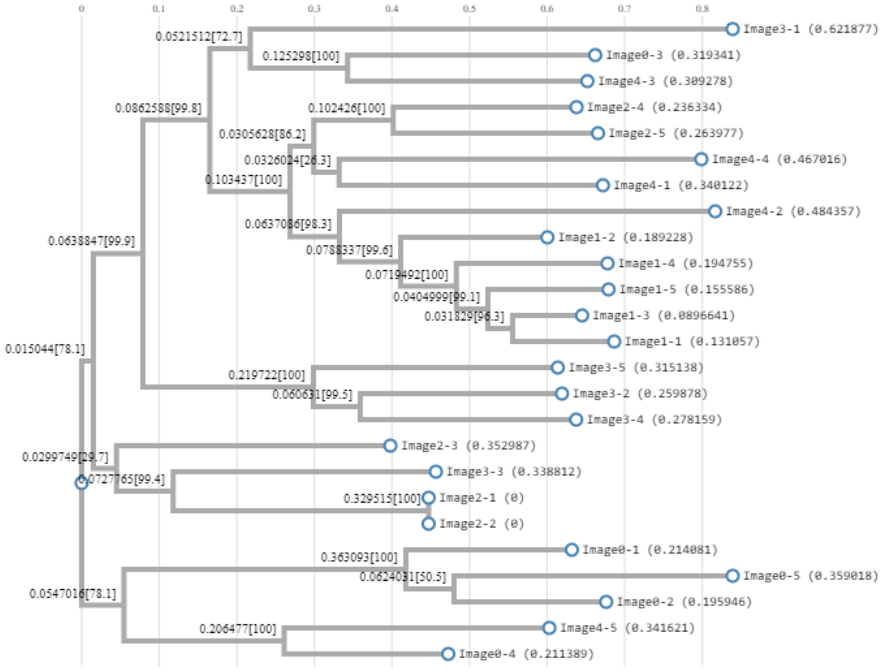

Fig. 10. Phylogenetic Results from the CLUSTALW Application

Visually, sequences from the same class tend to cluster within the same class or branch close to each other. For example, sequences from Class 0 can be grouped into a single large class, indicating that they are highly similar. Quantitatively, Fig. 10 shows the relatively short branch lengths between sequences of the same class. These short branch lengths indicated little change or difference between the sequences, indicating high similarity. The high bootstrap values in most branches indicated that the level of similarity shown by the tree was strongly supported by the data, making it reliable for further analysis.

If we relate Fig. 10 to the image object used (Fig. 9), we can analyze the similarity level of the resulting DNA sequences. Images 0 to 4 at the bottom of Fig. 10 show the DNA sequences obtained from MNIST image encoding, grouped according to their similarity. This indicates that sequences representing images with the same number (e.g., all 0 and 1 images) tend to cluster in the same branch or are close to each other. This indicates that the encoding process used to convert images into DNA sequences in this study successfully preserved the main visual features of the images, such that sequences derived from the same images (e.g., 0 and 1 images) were very similar.

Although this biological constraint is applied in encoding MNIST images into DNA sequences, for example, an image of number 0 will be translated into a specific nucleotide sequence pattern based on the visual characteristics of the lift, such as circular shapes or specific lines. Therefore, very similar images had similar nucleotide patterns reflected in the phylogenetic tree as close or very short branches (Fig. 10). Overall, the results of the phylogenetic analysis showed that the biological constraints in encoding succeeded in maintaining a close relationship between images with similar numbers. This proves that the approach to encoding MNIST images into DNA sequences effectively preserves the key visual characteristics that enable accurate clustering in the phylogenetic tree.

Additionally, there are weaknesses in encoding MNIST images into DNA sequences when applying biological constraints. The process of encoding an image into a DNA sequence naturally involves reducing information from its visual form to a nucleotide sequence representation. Therefore, some important image details may be lost or not accurately represented in the DNA sequence. This can cause sequences derived from visually similar images to be less identical at the DNA sequence level; therefore, branches on the phylogenetic tree may not fully reflect the original similarity of the images. The encoding process may not always accurately handle the variability of an image. For example, if there are slight variations in the image shape produced by noise or distortion, encoding may result in DNA sequences that differ from what they should be. For example, Image3-1 is in a different branch from Image3-2 to Image3-5, which could be due to the difference in line thickness between the number of images and other images of the same class. Thus, the resulting DNA sequences differed from the images in the same class.

This study has limitations, notably sensitivity to visual noise. Slight differences in pixel intensity due to noise or line thickness can cause significant variations in the encoded DNA sequence. In addition, there is a potential loss of visual information. The binary-to-DNA sequence mapping process is not a completely lossless representation. However, multiple sequence alignment (MSA) analysis and phylogenetic clustering are suitable approaches for evaluating information loss in the context of visual image encoding to DNA. Compared to other methods, MSA can resolve insertion-deletion-substitution errors without introducing logical redundancy or estimating any model parameters [38]. However, minor changes in the image structure can trigger significant differences at the sequence level, which causes the clustering accuracy to be imperfect (case example: Image3-1 vs Image3-2 through 3-5).

Furthermore, this study's approach is not intended for image compression and reconstruction but instead for the DNA representation of an image. Therefore, the compression ratio calculation is not relevant to this study. If this study focused on image reconstruction from DNA, metrics such as the Peak Signal-to-Noise Ratio (PSNR) or the Structural Similarity Index (SSIM) would be more suitable. However, since this study does not involve image decoding or reconstruction, these metrics cannot be applied directly in this study. Additionally, this study is still at the in silico encoding step; we have not physically synthesized the DNA, so the actual error rate is not yet available.

Nevertheless, outliers (as represented in Image 3-1) are examples of information partially lost due to differences in noise or line thickness, which is a natural manifestation of information loss. Therefore, the information loss identified through our MSA analysis is an acceptable trade-off appropriate for our intended target application. Therefore, this encoding method is more suitable for image representation applications in the DNA domain, where global visual feature preservation is more important than lossless image reconstruction. This research model is not recommended for applications requiring pixel-perfect visual information preservation (e.g., image reconstruction for medical applications or forensic imaging). This research model is not optimal for image representation, as it requires absolute information preservation.

3.6. Biological Constraints and Their Implications

Two significant biological constraints must be considered in DNA-based digital data encoding approaches: homopolymers and the GC content. Both constraints guarantee efficiency, data representation accuracy, and DNA sequence stability. Long sequences of similar nitrogenous bases, like AAAA or CCCC, are known as homopolymers. Technical issues could arise if homopolymers occur during the digital data encoding process. These include DNA synthesis and sequencing issues, as the enzymes involved often fail to identify repetitive sequences, increasing the possibility of sequencing errors. Homopolymers encountered during digital data encoding may cause incorrect DNA sequences, thus compromising the reliability of the stored data. To solve this problem, specific strategies are required in coding algorithms to reduce the occurrence of homopolymers with guaranteed base diversity in DNA sequences.

The GC content in a DNA sequence shows the proportion of cytosine (C) and guanine (G) bases. GC-content values, either high or low, affect DNA's thermal stability and structure. The three-hydrogen

bond between G-C makes high-GC sequences generally more stable upon heating, while the A-T pair only generates two hydrogen bonds. In contrast, excessive GC content reduces transcription efficiency or makes DNA synthesis and replication more difficult. The efficiency and consistency of DNA sequence rely on GC content in the digital data encoding process [39], [40]. Consequently, the encoding method must be specially designed to produce sequences with balanced GC content that are neither too high nor too low, thereby preventing disruption in the DNA synthesis and modification process.

Developing DNA-based digital data encoding techniques requires an understanding of and the ability to overcome biological constraints, such as homopolymers and GC content. Both determine the encoding and storage of data in DNA sequences, as well as the synthesis, sequencing, and manipulation of these sequences. DNA-based data storage systems may become more dependable and effective by utilizing attempts to overcome these limitations in the encoding process. These attempts would make it more valuable and applicable to various information technology and biotechnology applications. More dense and consistent data storage processes will become possible with a thorough understanding of these biological constraints and the implementation of appropriate solutions to overcome them.

This study utilizes the dataset provided by MNIST as a proof of concept. Nevertheless, there are many possible applications for this method of digital data encoding, such as digital archiving (proof of digital reports) [41], encryption to protect medical images (X-ray, CT-scan, MRI, etc.) [42]; secure image-based data filtering (image filtering systems can prevent unauthorized data dissemination) [43]; and cross-disciplinary bio-computing (integration of digital data with biological systems) [44]. Thus, this study opens up new opportunities for exploration at the intersection of computer science, data security, biotechnology and information management.

This study method differs from previous DNA encoding methods in terms of efficiency; although it is not the most efficient in terms of compression, it maintains the clarity of DNA sequences, which are easy to transcribe and translate. Regarding Stability, our study's homopolymer and GC-content constraints are strictly applied to ensure the sequence is stable for synthesis. In terms of Accuracy, by evaluating MSA and phylogenetic tree, this study revealed that the similarity between sequences is quite aligned with the visual similarity between images, even though there is unavoidable information loss. This is the first contribution to integrate transcriptional, translational, and MSA evaluation for images in the context of DNA-based data encoding.

3.7. Future Work and Application

Advances in DNA-based digital data encoding techniques offer a range of opportunities for research and application. These include improvements in the effectiveness and reliability of data storage in DNA by utilizing latent space-based deep learning technology. In addressing current biological constraints, this technology enables the development of data encoding and decoding models. This concept can influence how information is encoded and decoded in DNA sequences. This technology can be configured to identify complex data patterns and structures, enhance encoding efficiency, reduce homopolymer repetition, and maintain a balanced GC content. This method can improve resistance to errors during DNA synthesis and sequencing while generating the optimal DNA sequence for data storage by leveraging the complexity of neural networks.

Autoencoders are an artificial neural network that can generate more effective representations of input data and can be used to build this model. Using latent space, autoencoders can map digital inputs to DNA sequences more effectively and reliably. The decoder receives batch information, while the batch-free encoder can extract biological information [45]. Additionally, this model can be trained to consider biological constraints such as homopolymers and GC content, resulting in more stable and synthesizable DNA sequences.

The combination of DNA-based technology and encryption makes DNA an ideal candidate for use as a large-capacity, long-term, and stable data storage medium, making it suitable for data archiving [46]. Additionally, DNA can be used to store and analyze large volumes of genomic data in bioinformatics and medicine. Due to the complexity of encoding and decoding information in biological sequences, DNA can be used to store highly secure data. Deep learning models are utilized to develop DNA-based

encryption schemes, rendering them highly secure and resistant to decryption using traditional techniques.

Further research should investigate how entropy-aware coding techniques, such as context-adaptive symbol mapping or arithmetic coding, can minimize sequence length and optimize coding schemes to improve compression efficiency and reduce DNA synthesis costs. The latent space representations generated by deep neural network structures, such as variational or convolutional autoencoders, can be used to support these methods. This ensures that the encoded sequences retain their structural accuracy and biological viability compared to the original data. Additionally, this combination can enhance the effectiveness and reliability of data storage systems while addressing biological constraints. DNA could become a safe and effective way to store and manage data if this technology is researched and developed.

4. Conclusion

The object used in this study, the MNIST Digital Handwriting Dataset, was tested to determine whether the research method could be applied to more complex images. Based on the evaluation using multiple sequence alignment, the encoding results of the MNIST images had similar or close sequence characteristics, resulting in phylogenetic tree results that clustered closely together. However, a disadvantage of encoding images into DNA sequences is their sensitivity to the intensity of the original image. For thick and thin lines, a significant difference can cause the level of similarity to vary significantly across image classes. In addition to images, this encoding process can be applied to other types of digital data. However, if the digital data requires intact information or no loss is allowed, then the encoding method for the DNA Sequence cannot be used. This is because this technique results in the loss of data or information during the encoding of DNA sequences. This research demonstrates that the use of DNA for digital data storage has considerable potential, particularly in conjunction with deep learning technology. The findings of this study not only demonstrate success in overcoming biological barriers in DNA encoding but also provide a foundation for further research. The aim is to optimize DNA data storage to achieve a more efficient, secure, and sustainable data storage system in the future. This study represents the first step in developing a deep learning-based DNA data storage model utilizing the latent space. This deep learning model is intended to reduce and improve the DNA sequence produced after the basis for creating a DNA sequence free from biological constraints has been established. This strategy is expected to enhance the efficiency of DNA data storage and open new avenues for digital data storage technology.

Acknowledgement

The authors extend sincere thanks to the Department of Computer Science at IPB University, Sekolah Tinggi Teknologi Wastukancana, and Bunga Bangsa Foundation for their kind support and assistance with the research effort that led to this article. They were highly valued for their dedication to furthering computer science education and research.

Declarations

Author contribution. All the authors contributed equally to the main contributor to this work. All the authors have read and approved the final manuscript.
Funding statement. No funding was received for this study
Conflict of interest. The authors declare that there are no conflicts of interest associated with the research presented in this study.
Additional information. No additional information is available for this paper.

References

[1] A. Bessa-Silva, "Fasta2Structure: a user-friendly tool for converting multiple aligned FASTA files to STRUCTURE format," *BMC Bioinformatics*, vol. 25, no. 1, p. 73, Feb. 2024, doi: 10.1186/s12859-024-05697-7.

[2] X. Zhang and F. Zhou, "An Encoding Table Corresponding to ASCII Codes for DNA Data Storage and a New Error Correction Method HMSA," *IEEE Trans. Nanobioscience*, vol. 23, no. 2, pp. 344–354, Apr. 2024, doi: 10.1109/TNB.2024.3356522.

[3] C. Zhang *et al.*, "The Historical Evolution and Significance of Multiple Sequence Alignment in Molecular Structure and Function Prediction," *Biomolecules*, vol. 14, no. 12, p. 1531, Nov. 2024, doi: 10.3390/biom14121531.

[4] E. Upenik, D. Lazzarotto, M. Testolina, and T. Ebrahimi, "On the performance of learning-based image compression as source coding for JPEG DNA," in *Applications of Digital Image Processing XLVII*, Sep. 2024, vol. 13137, p. 31, doi: 10.1117/12.3031848.

[5] W. Wu, L. Xiang, Q. Liu, and K. Yang, "Deep Joint Source-Channel Coding for DNA Image Storage: A Novel Approach With Enhanced Error Resilience and Biological Constraint Optimization," *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, vol. 9, no. 4, pp. 461–471, Dec. 2023, doi: 10.1109/TMBMC.2023.3331579.

[6] I. Preuss, M. Rosenberg, Z. Yakhini, and L. Anavy, "Efficient DNA-based data storage using shortmer combinatorial encoding," *Sci. Rep.*, vol. 14, no. 1, p. 7731, Apr. 2024, doi: 10.1038/s41598-024-58386-z.

[7] P. M. Schwarz and B. Freisleben, "Data recovery methods for DNA storage based on fountain codes," *Comput. Struct. Biotechnol. J.*, vol. 23, pp. 1808–1823, Dec. 2024, doi: 10.1016/j.csbj.2024.04.048.

[8] D. Nachtigall Lazzarotto, J. Encinas Ramos, M. Testolina, and T. Ebrahimi, "Storing images and point clouds on DNA support with fountain codes," in *Applications of Digital Image Processing XLVII*, Sep. 2024, vol. 13137, p. 39, doi: 10.1117/12.3030612.

[9] L. Li, "Image encryption algorithm based on hyperchaos and DNA coding," *IET Image Process.*, vol. 18, no. 3, pp. 627–649, Feb. 2024, doi: 10.1049/ipr2.12974.

[10] Zeenath, K. DurgaDevi, and J. W. Carey M, "An Efficient Image Encryption Scheme for Medical Image Security," *Int. J. Electr. Electron. Res.*, vol. 12, no. 3, pp. 964–976, Aug. 2024, doi: 10.37391/ijeer.120330.

[11] T. Heinis, R. Sokolovskii, and J. J. Alnasir, "Survey of Information Encoding Techniques for DNA," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–30, Apr. 2024, doi: 10.1145/3626233.

[12] J. H. D. B. Gervasio, H. da Costa Oliveira, A. G. da Costa Martins, J. B. Pesquero, B. M. Verona, and N. N. P. Cerize, "How close are we to storing data in DNA?," *Trends Biotechnol.*, vol. 42, no. 2, pp. 156–167, Feb. 2024, doi: 10.1016/j.tibtech.2023.08.001.

[13] T. Buko, N. Tuczko, and T. Ishikawa, "DNA Data Storage," *BioTech*, vol. 12, no. 2, p. 44, Jun. 2023, doi: 10.3390/biotech12020044.

[14] S. Wang, X. Mao, F. Wang, X. Zuo, and C. Fan, "Data Storage Using DNA," *Adv. Mater.*, vol. 36, no. 6, p. 2307499, Feb. 2024, doi: 10.1002/adma.202307499.

[15] W. Alexan, E. Mamdouh, A. Aboshousha, Y. S. Alsahafi, M. Gabr, and K. M. Hosny, "Stegocrypt: A robust tri-stage spatial steganography algorithm using TLM encryption and DNA coding for securing digital images," *IET Image Process.*, vol. 18, no. 13, pp. 4189–4206, Nov. 2024, doi: 10.1049/ipr2.13242.

[16] B. Cao *et al.*, "Efficient data reconstruction: The bottleneck of large-scale application of DNA storage," *Cell Rep.*, vol. 43, no. 4, p. 113699, Apr. 2024, doi: 10.1016/j.celrep.2024.113699.

[17] K. O. Mohammed Aarif, V. Mohammed Yousuf Hasan, A. Alam, K. Shoukath Ali, and B. Pakruddin, "Decoding DNA: Deep learning's impact on genomic exploration," in *Deep Learning in Genetics and Genomics*, Elsevier, 2025, pp. 77–95, doi: 10.1016/B978-0-443-27574-6.00005-9.

[18] A. Usmani and L. Wiese, "DNA-Based Storage of RDF Graph Data: A Futuristic Approach to Data Analytics," *IEEE Access*, vol. 11, pp. 129931–129944, 2023, doi: 10.1109/ACCESS.2023.3332254.

[19] Yixun Wei, "Enlarge Practical DNA Storage Capacity: The Challenge and The Methodology," *University Of Minnesota*, pp. 1-24, 2023. [Online]. Available at: https://www.proquest.com/openview/551d0656f073ab423c2fb8f763c09470/1?pq-origsite=gscholar&cbl=18750&diss=y.

[20] L. Yunfei and Z. Xuncai, "Highly Robust DNA Data Storage Based on Controllable GC Content and homopolymer of 64-Element Coded Tables," *bioRxiv*. pp. 2023–2029, Sep. 29, 2023, doi: 10.1101/2023.09.27.559852.

[21] D. Landsman and K. Strauss, "The DNA Data Storage Model," *Computer (Long. Beach. Calif).*, vol. 56, no. 7, pp. 78–85, Jul. 2023, doi: 10.1109/MC.2023.3272188.

[22] X. Li, M. Chen, and H. Wu, "Multiple errors correction for position-limited DNA sequences with GC balance and no homopolymer for DNA-based data storage," *Brief. Bioinform.*, vol. 24, no. 1, pp. 1–11, Jan. 2023, doi: 10.1093/bib/bbac484.

[23] M. B. S. Al-Shuhaib and H. O. Hashim, "Mastering DNA chromatogram analysis in Sanger sequencing for reliable clinical analysis," *J. Genet. Eng. Biotechnol.*, vol. 21, no. 1, p. 115, Dec. 2023, doi: 10.1186/s43141-023-00587-6.

[24] Y. Liu, X. He, and X. Tang, "Capacity-Achieving Constrained Codes with GC-Content and Runlength Limits for DNA Storage," in *2022 IEEE International Symposium on Information Theory (ISIT)*, Jun. 2022, vol. 2022-June, pp. 198–203, doi: 10.1109/ISIT50566.2022.9834494.

[25] B. Pei *et al.*, "A Novel DNA-Based Dual-Mode Data Storage System with Interrelated Concise and Detailed Data," *Small Sci.*, vol. 4, no. 11, p. 2400094, Nov. 2024, doi: 10.1002/smsc.202400094.

[26] S. Jo, H. Shin, S. Joe, D. Baek, C. Park, and H. Chun, "Recent progress in DNA data storage based on high-throughput DNA synthesis," *Biomed. Eng. Lett.*, vol. 14, no. 5, pp. 993–1009, Sep. 2024, doi: 10.1007/s13534-024-00386-z.

[27] Q. Huang *et al.*, "Emerging preservation materials for long-term DNA-based data storage," *Chem. Eng. J.*, vol. 509, p. 161245, Apr. 2025, doi: 10.1016/j.cej.2025.161245.

[28] A. Doricchi *et al.*, "Emerging Approaches to DNA Data Storage: Challenges and Prospects," *ACS Nano*, vol. 16, no. 11, pp. 17552–17571, Nov. 2022, doi: 10.1021/acsnano.2c06748.

[29] Y. Zheng, B. Cao, X. Zhang, S. Cui, B. Wang, and Q. Zhang, "DNA-QLC: an efficient and reliable image encoding scheme for DNA storage," *BMC Genomics*, vol. 25, no. 1, p. 266, Mar. 2024, doi: 10.1186/s12864-024-10178-5.

[30] H. Du, S. Zhou, W. Yan, and S. Wang, "Study on DNA Storage Encoding Based IAOA under Innovation Constraints," *Curr. Issues Mol. Biol.*, vol. 45, no. 4, pp. 3573–3590, Apr. 2023, doi: 10.3390/cimb45040233.

[31] A. Rasool, J. Hong, Q. Jiang, H. Chen, and Q. Qu, "BO-DNA: Biologically optimized encoding model for a highly-reliable DNA data storage," *Comput. Biol. Med.*, vol. 165, p. 107404, Oct. 2023, doi: 10.1016/j.compbiomed.2023.107404.

[32] X. Zhang, B. Qi, and Y. Niu, "A dual-rule encoding DNA storage system using chaotic mapping to control GC content," *Bioinformatics*, vol. 40, no. 3, Mar. 2024, doi: 10.1093/bioinformatics/btae113.

[33] A. A. Yassin, A. Mohammed Rashid, A. J. Yassin, and H. Alasadi, "A novel image encryption scheme based on DCT transform and DNA sequence," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 21, no. 3, p. 1455, Mar. 2021, doi: 10.11591/ijeecs.v21.i3.pp1455-1464.

[34] M. Courel *et al.*, "GC content shapes mRNA storage and decay in human cells," *Elife*, vol. 8, pp. 1–32, Dec. 2019, doi: 10.7554/eLife.49708.

[35] C.-A. Canfield and P. C. Bradshaw, "Amino acids in the regulation of aging and aging-related diseases," *Transl. Med. Aging*, vol. 3, pp. 70–89, Jan. 2019, doi: 10.1016/j.tma.2019.09.001.

[36] X. Fang *et al.*, "A method for multiple-sequence-alignment-free protein structure prediction using a protein language model," *Nat. Mach. Intell.*, vol. 5, no. 10, pp. 1087–1096, Oct. 2023, doi: 10.1038/s42256-023-00721-6.

[37] D. Huo, D. Zhou, S. Yuan, S. Yi, L. Zhang, and X. Zhou, "Image encryption using exclusive-OR with DNA complementary rules and double random phase encoding," *Phys. Lett. A*, vol. 383, no. 9, pp. 915–922, Feb. 2019, doi: 10.1016/j.physleta.2018.12.011.

[38] R. Xie, X. Zan, L. Chu, Y. Su, P. Xu, and W. Liu, "Study of the error correction capability of multiple sequence alignment algorithm (MAFFT) in DNA storage," *BMC Bioinformatics*, vol. 24, no. 1, p. 111, Mar. 2023, doi: 10.1186/s12859-023-05237-9.

[39] M. H. Raza, S. Desai, S. Aravamudhan, and R. Zadegan, "An outlook on the current challenges and opportunities in DNA data storage," *Biotechnol. Adv.*, vol. 66, p. 108155, Sep. 2023, doi: 10.1016/j.biotechadv.2023.108155.

[40] Y. Cevallos *et al.*, "A brief review on DNA storage, compression, and digitalization," *Nano Commun. Netw.*, vol. 31, p. 100391, Mar. 2022, doi: 10.1016/j.nancom.2021.100391.

[41] J. McLeod and E. Lomas, "Record DNA: reconceptualising digital records as the future evidence base," *Arch. Sci.*, vol. 23, no. 3, pp. 411–446, Sep. 2023, doi: 10.1007/s10502-023-09414-w.

[42] B. Ahuja, R. Doriya, S. Salunke, M. F. Hashmi, and A. Gupta, "Advanced 5D logistic and DNA encoding for medical images," *Imaging Sci. J.*, vol. 71, no. 2, pp. 142–160, Feb. 2023, doi: 10.1080/13682199.2023.2178097.

[43] K. Cho and H. Bahn, "Evaluating Image DNA Techniques for Filtering Unauthorized Content in Large-Scale Social Platforms," *Appl. Sci.*, vol. 15, no. 8, p. 4539, Apr. 2025, doi: 10.3390/app15084539.

[44] S. Jia, H. Lv, Q. Li, C. Fan, and F. Wang, "DNA-based biocomputing circuits and their biomedical applications," *Nat. Rev. Bioeng.*, vol. 3, no. 7, pp. 535–548, Apr. 2025, doi: 10.1038/s44222-025-00303-8.

[45] Y. Liu, Z. Li, X. Chen, X. Cui, Z. Gao, and R. Jiang, "INSTINCT: Multi-sample integration of spatial chromatin accessibility sequencing data via stochastic domain translation," *Nat. Commun.*, vol. 16, no. 1, p. 1247, Feb. 2025, doi: 10.1038/s41467-025-56535-0.

[46] Y. Zhou, K. Bi, Q. Ge, and Z. Lu, "Advances and Challenges in Random Access Techniques for In Vitro DNA Data Storage," *ACS Appl. Mater. Interfaces*, vol. 16, no. 33, pp. 43102–43113, Aug. 2024, doi: 10.1021/acsami.4c07235.