RESEARCH ARTICLE

RFS-codec: A Novel Encoding Approach to Store Image Data in DNA

Abdur Rasool[1,*]

[1] Department of Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI 96822, United States

Abstract

DNA data storage is a promising technology that utilizes computer simulation and offers high-density and durable digital information storage. It is challenging to store massive image data in a small amount of DNA without losing the original data since nonspecific hybridization errors occur frequently and severely affect the durability of stored data. This work proposes a novel approach (RFS-codec) comprising an image fraction strategy and an innovative codec method to split and encode image data into DNA storage, respectively. The fraction strategy contributes by delivering a cost-effective solution for image storage in DNA. The codec method offers an encryption mechanism to convert binary data into DNA bases by avoiding hybridization errors and satisfying the critical bio-coding constraints responsible for DNA storage durability. The robustness of RFS-codec is computed with GC and homopolymer constraints. Experimentally, different image data are efficiently encoded and decoded successfully with 1.8 bit/nt average density. RFS-codec's results demonstrate substantial advantages in constructing cost-effective, scalable, and durable DNA data storage.

Keywords: DNA data storage, image fraction, codec approach, bio-coding constraints.

1 Introduction

The utilization and storage of massive data pose both conveniences and challenges. Consequently, effectively managing and preserving extensive data volumes has become a complex issue for data scientists. As projected by the International Data Corporation, the global data capacity is estimated to reach 175 ZB by 2025. However, current storage systems encounter obstacles concerning exorbitant expenses, substantial energy consumption, and low density. To address these concerns, Deoxyribonucleic Acid (DNA) has garnered significant attention from researchers due to its capacity for high parallelism, cost-effective maintenance, and substantial storage potential. DNA offers various unique advantages, including long-term stability lasting decades or centuries, in contrast to traditional media that require frequent updates. Furthermore, DNA allows easy replication using molecular biology techniques to prevent degradation [1–5].

In the late 20th century, Joe Davis pioneered the utilization of bacteria as a storage medium for abiotic

Submitted: 15 May 2025
Accepted: 22 June 2025
Published: 30 June 2025

Vol. 1, **No.** 1, 2025.
10.62762/JAIB.2025.146324

***Corresponding author:**
✉ Abdur Rasool
abdur@hawaii.edu

Citation
Rasool, A. (2025). RFS-codec: A Novel Encoding Approach to Store Image Data in DNA. *Journal of Artificial Intelligence in Bioinformatics*, 1(1), 41–50.

information, marking the initial endeavors to employ DNA for this purpose [6]. Bancroft et al. [7] made strides in 2001 by encoding excerpts from 'A Tale of Two Cities' into a DNA molecule, utilizing a codon-based approach analogous to encoding protein sequences. Subsequently, in 2012, Church et al. [1] encoded various types of data, including text and images, into corresponding DNA sequences. These sequences were stored within DNA and capable of accommodating errors arising from DNA sequencing and synthesis. Addressing the challenges associated with DNA synthesis, amplification, and sequencing, Song et al. [4] developed the de Bruijn graph and greedy path search-based strand assembly algorithm (DBGPS) in 2022. Notably, the algorithm successfully recovered 6.8 MB of image data from a severely corrupted sample, achieving a logical density of 1.30 bits per cycle.

Apart from these significances, image data usually undergoes compression before storage, and a mere mismatch can trigger extensive error propagation during decompression, resulting in distorted reproductions [4, 8]. Achieving such fidelity poses significant challenges for conventional DNA-based data storage systems, primarily due to the inherently stochastic nature of the PCR, sequencing, and rewriting processes [9]. Similarly, one of the crucial challenges of DNA data storage is the DNA synthesizing and sequencing cost. According to a recent report, the cost of storing 1MB of data using 100% chemical yield is approximately 1e-8 US$ [9]. In 2012, Church [1] reported 11,650 US$ for 0.65 MB; in 2020, Philipp [10] stated 530 US$ for 0.1 MB of data. Additionally, The existing DNA data storage studies [1, 4, 8] on image data only consider the specific image format; for example, Church [1] and DBGPS [4] provide the encoding-decoding for only JPG format. However, scientific data, specifically extinct animals image, have not been encoded/decoded into DNA storage.

Here, this study design and experimentally test a novel codec approach for a data storage system termed RFS-codec by addressing the issue of nonspecific error propagation during hybridization to provide durable DNA sequences with minimum cost. Acronyms of the proposed RFS-codec approach are Rasool's Fraction-Strategy for encoding image data in DNA. This approach offers a novel image fraction strategy that divides the image into 16 equal parts to combat the cost issue of DNA sequencing. Moreover, the proposed approach develops an innovative transcoding mechanism to convert the digital image into binary data and then map it into DNA bases for storage. This mechanism delivers a high-density scalable and robust mapping by controlling the bio-coding constraints (GC-content (GC) and homopolymer (HP)) and avoiding propagation of hybridization error. Experimentally, RFS-codec is applied on different extinct animals' image due to; (1) preserve the genetic disorder information for centuries for the next generation, (2) such data stored in existing mediums would not be reliable to analyze after 15 30 years; the stored information in DNA can be accessed and analyzed accurately to differentiate the health record of forefathers after decades. Thus, the significant contributions of this study are:

- A novel RFS-codec approach is proposed to tackle the nonspecific error propagation issue by reporting a durable, scalable, highly dense, and cost-effective encoding mechanism to store image data in DNA.
- The proposed image fraction strategy combats the DNA sequencing cost issue, and novel encoding control the errors by satisfying the two crucial bio-coding constraints.
- The robustness of the proposed codec is assessed by encoding image data, and results demonstrated a substantial performance than prior works.

The rest of the work is structured as follows: Section 2 presents the literature, Section 3 describes the Preliminaries, Section 4 proposes the RFS-codec, Section 5 reports the experiments and results evaluation, and the conclusion is in Section 6.

2 Literature Review

DNA computing has gained significance in image encryption due to its notable attributes of extensive parallelism and remarkable information density. This computing approach encompasses biological operations and algebraic manipulations performed on DNA sequences, such as DNA addition, subtraction, and XOR operations. In image encryption, DNA-based encryption algorithms utilize DNA encoding and DNA computing techniques [4, 8, 11, 12]. Images hold great importance as a prevalent data type extensively utilized and stored in cloud storage and social media platforms. Image-based data possesses two distinct characteristics. Firstly, it exhibits a substantial volume owing to its exponential growth and generation through social media platforms like Facebook, Snapchat, and Instagram. Consequently,

this voluminous nature escalates the demand for high-capacity storage systems. Secondly, image-based data necessitates fault tolerance. Numerous previous studies have explored the approximate storage of images, employing technologies such as flash memory or non-volatile memory to reduce storage overhead or enhance image robustness [13, 14]. As such, these two attributes of images align well with the error-prone yet densely packed nature of DNA storage systems. However, prior investigations have not explicitly focused on the robustness of images within DNA storage. While some existing studies have successfully implemented image-based binary data in DNA storage, their primary emphasis lies in assessing the feasibility of DNA storage for various data types, including images [1, 15].

The existing literature on DNA-based image encryption and storage is limited in its specific focus [4, 8, 11, 12]. For instance, a novel approach has been proposed for archiving images in DNA, wherein missing and erroneous oligos are rectified using specialized learning methods instead of costly coding redundancy. This approach incorporates a quantization scheme that reduces the image color palette to 8 intensity levels per channel and employs Hilbert-space filling curves, differential coding, and Huffman coding for intensity level compression [8]. However, it should be noted that such compression techniques may result in significant error propagation when confronted with missing or mismatched oligos. Another study introduced an algorithm for DNA sequence-based image encryption, primarily targeting large images [12]. This approach utilizes the first DNA sequence to generate a scrambling sequence for pixel scrambling and the second to generate three DNA templates for pixel replacement. Notably, the authors do not employ DNA biological operations in their image encryption methodology, and their focus primarily revolves around evaluating storage density and error correction. Hence, the efficient storage of images in DNA using cost-effective and durable storage mediums emerges as a critical and intriguing research concern.

3 Preliminaries

This section introduces the related techniques which are pre-established and reported by other authors.

3.1 GC Content

GC content can influence the stability of double-stranded DNA, as G-C pairs have three hydrogen bonds, while A-T pairs have two, making G-C-rich DNA more stable. Identify functional elements such as coding regions, regulatory regions, and repetitive elements, as they often have distinct GC content compared to the surrounding DNA. GC content refers to the percentage of Guanine (G) and Cytosine (C) nucleotides in a DNA molecule, compared to the total number of nucleotides (Adenine (A) and Thymine (T)), as given in (1). It measures the proportion of G-C pairs in a DNA sequence and is used as an indicator of DNA's structural and functional characteristics [15].

$$GC\ content\ (L) = \frac{|G| + |C|}{|A| + |T| + |G| + |C|} \times 100\% \quad (1)$$

3.2 Homopolymer

Homopolymer (HP) DNA coding constraints refer to the limitations on the sequencing of DNA that contain consecutive nucleotides of the same type (e.g., AAAAAA or TTTTTT). These HP regions pose challenges for DNA sequencing technology, as it can be difficult to accurately determine the homopolymer tract length due to base calling errors [2].

Apart from the mathematical definition of fraction (numerator/denominator), the fraction definition is optimized for the proposed image fraction strategy. When an object (image) is divided into several equal parts, then each part (sub-image) is called a fraction. Facilitate the storage of digital information within nucleotides demands defining the DNA data density concept with exponential growth factor $\hat{Y}_{bn}$ [16].

Definition 1. *The highest quantity of digital data bits* (b) *accommodated per base or nucleotide* (nt) *is called as data density of DNA data storage* (D_{bn}).

$$D_{bn} = \lim_{n\to\infty} \left(\frac{1}{n} \log_2 \xi_{bn} \right) = \log_2 \hat{Y}_{bn} \left(\frac{b}{nt} \right) \quad (2)$$

where ξ_{bn} *is a DNA library that has information of data bits and nucleotides.*

Moreover, Theorem 1 is utilized by taking the specific code-length $(n-1)$ for the production of the optimal DNA codewords. Additionally, Theorem 2 is derived with a particular term of $(d-1)$ Hamming distance with $\mathcal{L}$ to generate codes of DNA that must meet the crucial bio-coding constraints [16].

Theorem 1. *A DNA coding set of* n *length will be less than a codeword of* $n-1$ *length for a minimum Hamming distance of* $(0 \leq d \leq n)$, *while GC content* $(0 < L < n)$.

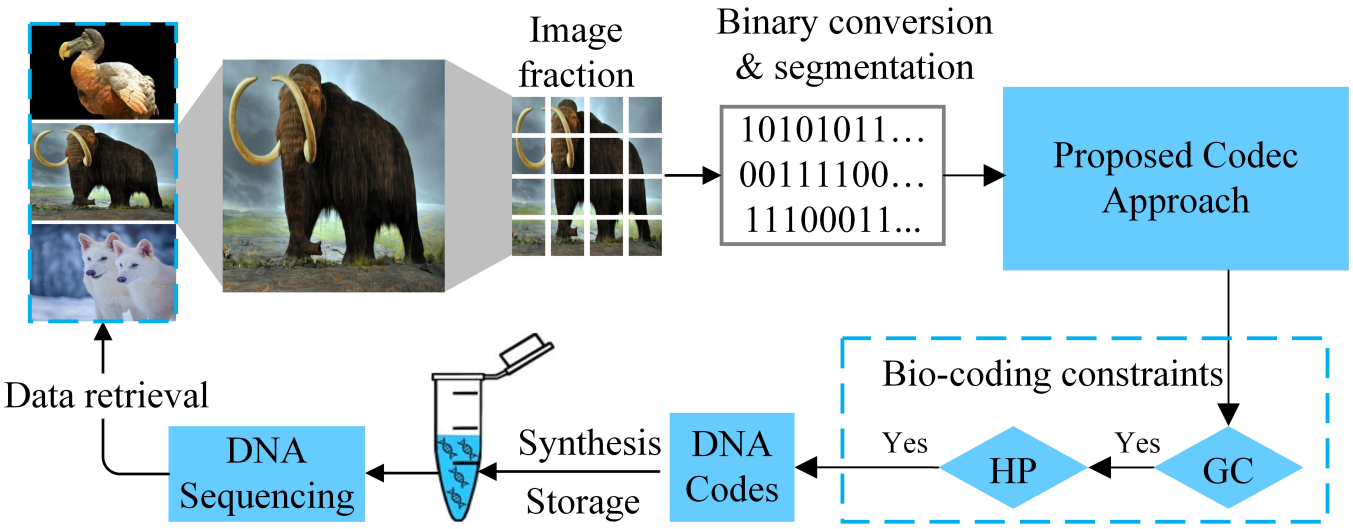

Figure 1. The workflow diagram of the proposed RFS-codec approach.

Proof. In equation (3), the DNA sequence β_i with β_1 codeword and length n of GC content L, there will be position j in which

$$A_4^{GC}(n,d,L) \leq \left\lfloor \frac{2n}{\omega} A_{GC4}(n-1,d,L-1) \right\rfloor \quad (3)$$

$\left\lfloor \frac{L\alpha_1}{2n} \right\rfloor$ DNA codes have the next DNA base C, or, in a given particular position, this code can be G. Besides, the mean GC content will be minimum than the original L. Similarly, considering the DNA codewords with deleting position j, it can produce $n-1$ and $L-1$ DNA codewords with minimum distance. Thus, equation (4) is parallel, but it differs with L for different positions where

$$A_4^{GC}(n,d,L) \leq \left\lfloor \frac{2n}{n-L} A_{GC4}(n-1,d,L) \right\rfloor \quad (4)$$

$\left\lfloor \frac{(n-L)\alpha_1}{2n} \right\rfloor$ reports A's or T's. □

In this work, the variations in (3) and (4) brought to receive the DNA code's upper bounds on $A_{GC4}(n,d,L)$ with $n=d$, $n=L$, or $L=0$ positions. Meanwhile, other coding bounds (i.e., lower bounds) can also be achieved by changing different variables; for example, if code length n is constant ($n=d$), then (3) can consider with $n=L$ and (4) with $L=0$.

Theorem 2. *A DNA sequence with the highest code-length n and lowest distance $(d-1)$ can satisfy the GC content L in the following lower bounds inequality:*

$$A_4^{GC}(n,d,L) \geq \frac{\binom{n}{L}2^n}{\sum_{r=0}^{d-1}\sum_{i=0}^{\min\{\lfloor \frac{r}{2} \rfloor, L, n-L\}} \binom{L}{i}\binom{n-L}{i}\binom{n-L}{r-2i}2^{2i}} \quad (5)$$

Proof. The numerator $\binom{n}{L}2^n$ of (5) presents the total DNA codewords in a sequence β_i with GC-content L. In contrast, the denominator provides the DNA codes with Hamming distance $d-1$ for a sequence β_i, while $\binom{L}{i}\binom{n-L}{i}\binom{n-L}{r-2i}2^{2i}$ indicates the lower bounds. It yields the DNA codes for a sequence β with L that must meet the Hamming distance d to avoid the error r from the entire DNA sequence. □

4 Proposed Approach

This paper proposed a novel fraction strategy to split images and a new codec method to store image data in DNA by reducing the cost and errors.

In the following sub-sections, this work introduces the novelty of image fraction strategy and codec method as significant contributions. Moreover, it employed bio-coding constraints, briefly presented in Section III. Meanwhile, DNA synthesis, storage, and data retrieval (sequencing) have been concisely described due to the neglect of wet lab experiments in this work. The schematic diagram of the proposed approach for DNA-based data storage is illustrated in Figure 1.

4.1 Image Fraction Strategy

A novel strategy is introduced to store an image by splitting it into smaller sub-images as fractions. The purpose of presenting this strategy is to provide an alternative method of random access and to facilitate

the analysis and organization of stored data in DNA, offering a cost-effective solution and advantages from the following perspectives.

1. State-of-the-art encoding models [1, 5, 15] have utilized the rotating code technique to adhere to biochemical constraints. However, if nucleotide changes occur within the DNA sequence strands, error propagation affects subsequent subsequences, leading to incorrect data decoding. Thus, the fragmentation of images can effectively mitigate such error propagation, ensuring accurate decoding.

2. This fragmentation strategy allows selective retrieval of image chunks instead of complete images, significantly reducing costs and time requirements. Practitioners can conveniently identify relevant patterns (pertinent to genetic traits or diseases) from specific image portions, obviating the need to access the entire image.

3. Since DNA storage stores lengthy sequences, accessing and manipulating them can pose challenges; this strategy optimizes storage, organizes data, and facilitates structured indexing, simplifying the retrieval of specific information when required.

4. Additionally, it can reduce computational complexity and computing memory.

The strategy splits an image into smaller segments based on the specified number of rows (*split_h*) and columns (*split_w*), using the Python Imaging Library (PIL), and OS and math modules. A Fraction algorithm (Algorithm 1) is proposed to determine the number of rows (*split_h*) and columns (*split_w*) parameters to split an image by identifying input file format and size.

To calculate the dimensions of the image segments and the remaining width ($w = img\_size$) and height ($h = img\_size$), this strategy uses nested for loops to traverse the grid in row-major order. The minimum column_height (C_h) (6) and row_width (R_w) (7) are computed as follows:

$$C_h = \left\lceil \frac{h}{split\_h} \right\rceil \tag{6}$$

$$R_w = \left\lceil \frac{w}{split\_w} \right\rceil \tag{7}$$

However, the last column_height C_{hl} and last row_width R_{hl} can be calculated as follows:

$$C_{hl} = h - (3 \cdot C_h) \tag{8}$$

$$R_{hl} = h - (3 \cdot R_w) \tag{9}$$

The fraction algorithm is developed with four conditions inside the nested loops by iterating through a 2D grid that determines how the image should be cropped.

- If the current grid cell is not on the last row R_{wl} and not on the last column C_{hl}, the image is cropped with dimensions C_h and R_w. If the current grid cell is on the last row R_{wl} but not the last column C_{hl}, the image is cropped with dimensions R_w and C_{hl}.
- If the current grid cell is on the last column C_{hl} but not the last row R_{wl}, the image is cropped with dimensions R_w and C_h. If the current grid cell is on the last row R_{wl} and last column C_{hl}, the image is cropped with dimensions R_{wl} and C_{hl}.

The outer loop iterates through the rows using index i, and the inner loop iterates through the columns using index j. Depending on the position of the current grid cell (i, j), the fraction algorithm crops the input image using appropriate dimensions to crop a segment.

4.2 Codec Method

The inspiration for the proposed codec method is based on Fountain encoding [2], which serves as the foundational scheme. In the Fountain method, this mapping is applied after the XORed function, and basic coding constraints (GC and HP) are used to judge sequence satisfaction criteria. The sequences that do not satisfy these constraints are directly discarded, which loses the portion of binary segments and causes the original information reliability. Therefore, we have designed a new codec method to avoid data drops after the basic mapping and avoid the errors significantly.

The binary segments are read through the encoder based on the proposed method distributed in Tables 1 and 2. According to the number of base repeats in the DNA sequence, tandem repeats are divided into single and two base types. A step-by-step functionality of each Table and the base type is elaborated as follows.

1. The single-base repeat sequences are unstable in organisms and often mutate in replication

Algorithm 1 Fraction algorithm to split an image into 16 equal sub-images

Input: Function `cut` with `split_h`, `split_w` parameters, Image width w = `img_size`[0] and height h = `img_size`[1], min_height (C_h), min_width (R_w), last row width (R_{wl}) and last column height (C_{hl})
Output: Split an image into 16 chunks

```
Initialize count variable to 0
Get image_size (w and h)
for i = 0 to split_h - 1 do
    for j = 0 to split_w - 1 do
        if not last row and not last column then
            crop image using R_w and C_h
        else if last row and not last column then
            image using R_w and C_hl
        else if not last row and last column then
            crop image using R_wl and C_h
        else if last row and last column then
            image using R_wl and C_hl
        end if
    end for
end for
Calculate C_h and R_w using (6) and (7), respectively
Calculate C_hl and R_wl using (8) and (9) for edge segments
return Generate smaller sub-images into 16 parts
```

and transcription. DNA fragments with single base repeats are prone to base mutation and base deletion. In DNA data storage, copying and amplifying DNA sequences is necessary. In order to realize the durable storage of DNA sequences and avoid initial codons, this paper constructs a single base repeat sequence base library $S_1 \in \{\text{AAA}, \text{TTT}, \text{CCC}, \text{GGG}\}$ and $S_2 \in \{\text{ATG}, \text{TAA}, \text{TGA}, \text{TAG}, \text{TGG}\}$ to avoid the above base sequences when performing encoding. This mapping is performed in three stages; first 4 bits, intermediate bits Int_b, and the last bit or 2 bits l_b.

2. After converting the first four binary bits to DNA bases by utilizing the Fountain encoding scheme, it is necessary to avoid the base sequence binding with the single base repeat in the base library of DNA. When the current base pair is converted typically, it is essential to construct a two-bit base set, $B = \{\text{AT}, \text{TA}, \text{TG}, \text{TT}, \text{AA}, \text{GG}, \text{CC}\}$, given in Table 1. The encoder reads the next binary bit from the Int_b, corresponding to the prior base pair, and maps it with a single base. It continually scans whether the base formed by each generated base and the previous base is not repeated. The base conversion of the DNA sequence shall be carried out accordingly.

3. Meanwhile, if the next 1 binary bit, with a prior base pair, does not exist in set B, the encoder considers the two consecutive binary bits and maps accordingly. However, if the previous base pair with the next 2 bits do not exist in set B, but the corresponding bits are in X; then map with a single base. In order to generate durable DNA codes by avoiding two base tandem repeats, some binding sites are left empty with the "-" mark in Table 1.

4. The encoder is trained to read the l_b, or two bits shall be mapped with two bases by utilizing Table 2.

Table 1. Mapping of intermediate bits to nucleotides.

	AT	TA	TG	TT	AA	GG	CC	X
A	0	-	-	0	-	0	0	00
T	10	1	1	-	10	10	10	01
C	11	0	0	10	0	11	-	10
G	-	-	-	11	11	-	11	11

Table 2. Mapping Of The last one bit or 2 bits to a pair of nucleotides.

0	1	00	01	10	11
AC	TC	CG	GA	GT	GC

The construction of optimal DNA sequences through the *codec* method is structured in Algorithm II. In order to determine the position of each DNA sequence fragment and meet the requirements of random reading, two reserved marker bases (RB) are used to represent the information of sequence number in the DNA sequence segment, i.e., `ATTATT` represents the fragment number (F#) "0", `ATTAAG` indicates the F# "1", etc. The corresponding relationship between the RB and the F# is given in Table 3. After assigning the index to each DNA sequence related to each data, eight DNA fragments are taken into account from each data. This helps to achieve an orderly reading of DNA sequence segments stored in DNA data. When DNA data storage is carried out to restore DNA sequence files to the computer, each stored data can be quickly located by the F# represented by the RB information.

4.3 Storage and Retrieval

This stage belongs to the wet lab experiments with biotechnology. This paper briefly describes the

Algorithm 2 Pseudocode of proposed codec method

Input: Mapping of first four bits b_4, intermediate bits Int_b, and last bit l_b w.r.t Tables 1 or 2 (T1, T2), excluding set S and prior 2 bases $2B$

Output: DNA sequences

```
mapping initialization
convert first 4 bits b_4 as per Fountain
while b_5 ≤ Int_b < l_b do                ▷ mapping Int_b
    if (2B, Int_b) in T1 then
        map Int_b w.r.t prior 2 bases
    else
        read two consecutive Int_b
        if 2B ∉ T1 and Int_b ∈ S then
            map two bits w.r.t 2B with X of T1
        else
            map two Int_b w.r.t prior 2 bases
        end if
    end if
end while
for Int_b < l_b do
    for Int_b ≥ l_{b+1} do
        encode the last bit or two bits as per T2
    end for
end for
Output: DNA sequences
```

Table 3. DNA sequence fragment number and corresponding reserved base.

F#	0	1	2	3
RB	ATTATT	ATTAAG	ATTCCG	AAGATT
F#	4	5	6	7
RB	AAGAAG	AAGCCG	CCGATT	CCGAAG

DNA synthesis to store digital information and the sequencing process to retrieve the original data. The proposed codec method produces a DNA sequence file comprising strings representing a set of DNA sequences, which a DNA array synthesizer can interpret. Several commercial suppliers are available to manufacture the DNA molecules. Illumina's DNA sequencing is used for its wide adoption and minimal read error rates. Illumina offers a range of benchtop sequencing platforms, such as the NextSeq Series, providing 120 GB output and 400 million sequence reads. Predefined sequencing adaptors must be added to both ends of the DNA molecules to compatibility with Illumina machines. This enables the sequencing of a file containing a minimum of 200 kB of digital information at a consumable cost of approximately US$600, requiring 3 hours of hands-on operating time and 17 hours of measurement time. The decoder reads either a text file or a FASTQ file provided by the sequencer. It reverses the encoding procedure by initially mapping the sequences back to bits and decrypting them to retrieve the original information.

5 Experiments And Results

5.1 Implementation Process

Extinct animals, e.g., woolly mammoth[1], dire wolf[2], and pink pigeon[3]. However, any other image can also be considered. This data is processed by the proposed image fraction strategy to split it into 16 equal parts. Then each sub-image part is converted into binary bits using a Python function *TransBin*. The binary data is mapped according to the proposed novel *Codec* method by importing Python packages (`codecs`) to generate the premier DNA sequences. These sequences were assessed through a strict criterion to construct reliable DNA codes that satisfy the bio-coding constraints (GC and HP). For instance, the control for GC was set to 40–60% and HP ≤ 4. The satisfaction of these two crucial constraints leads to avoiding the propagation of nonspecific hybridization errors in the optimal DNA sequences. As this study aims to generate cost-effective, durable, high-density DNA codes, the wet-lab experiment (DNA synthesis, storage, and sequencing) is not considered. Moreover, the optimal DNA codes are decoded using the reverse process of the proposed *Codec* method, and images are retrieved successfully.

5.2 Constraints Satisfaction

To confirm the DNA code satisfaction with bio-constraints, 100 random DNA sequences of image files are evaluated, providing the balanced GC content with the proposed codec method. Figure 2 demonstrates the comparison of unconstrained (a) and constrained (b) sequences without and with the proposed RFS-codec approach, respectively. The constraint-based DNA sequences ensure the reduction of nonspecific hybridization errors, which cause DNA durability.

5.3 Encoding Efficiencies

Experimentally three different image (img) files with different sizes are encoded to assess the proposed approach efficiencies. Table 4 reports the RFS-codec's

[1] https://www.popsci.com/woolly-mammoth-dna-brought-life-elephant-cells/

[2] https://news.ucsc.edu/2025/04/dire-wolf-genome/

[3] https://www.newscientist.com/article/mg26234950-500-how-dodo-de-extinction-is-helping-rescue-the-extraordinary-pink-pigeon/

Table 4. Performance comparison of the RFS-codec approach with different files on various factors.

File	Size (MB)	Total DNA nt	HP	GC	Time (s)	D_{bn}
img1	3.67	17056475	3	49	247.31	1.8
img2	8.63	40015709	2	50	561.48	1.81
img3	7.22	33539894	3	54	474.07	1.8

Table 5. The comparative analysis of the proposed RFS-codec approach with prior studies.

Author	Year-Refs.	Coding method	Bio-constraints	D_{bn}
Church	2012-[1]	1bit to 1base	HP	0.83
Erlich	2017-[2]	Fountain	GC	1.57
Organick	2018-[15]	RS coding	GC content, no-runlength	1.1
Song	2022-[4]	De Bruijn graph	GC, Melting temperature	1.30
Rasool	**2023**	**RFS-codec**	**GC, HP**	**1.8**

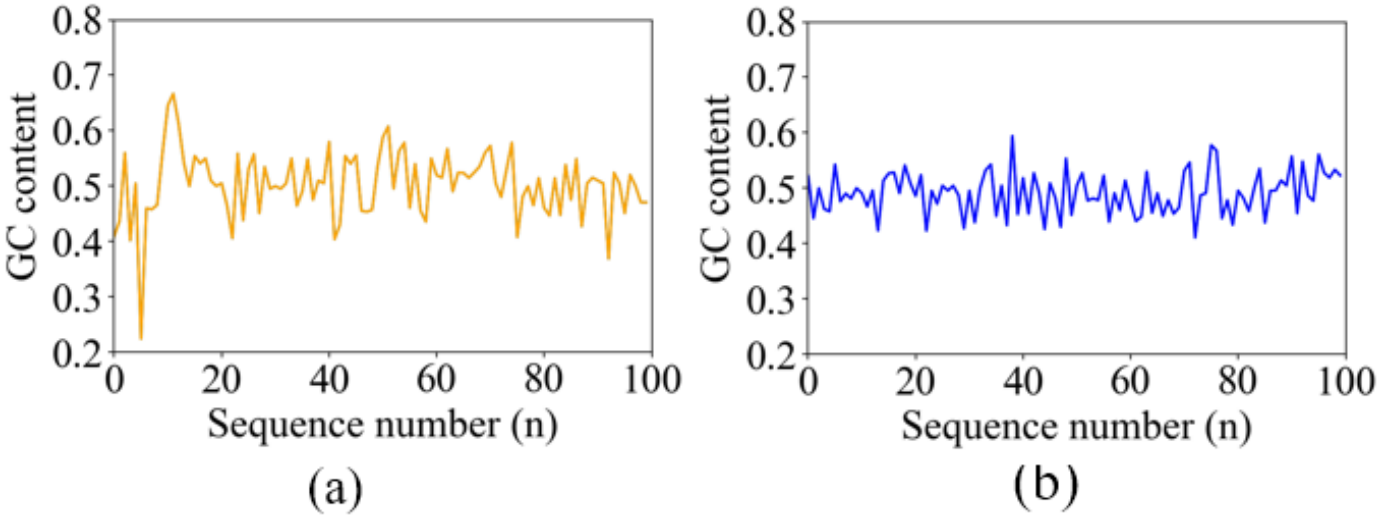

Figure 2. The comparison of GC content satisfaction (a) without and (b) with the proposed RFS-codec approach.

performance with different parameters, i.e., total nucleotides (nt), the rate of bio-constraints (HP and GC), and each file's encoding and decoding time. All files are 100% decoded with different densities and times. The satisfaction of constraints, running time, and adequate density (D_{bn}) reports the effectiveness of the proposed RFS-codec approach in terms of durable and high-density DNA data storage.

5.4 Cost-effectiveness

This study introduced an image fraction strategy for DNA storage of images, showcasing its cost-efficient data storage and retrieval capabilities. A comparative analysis against Goldman's (2013) work reveals a remarkable cost advantage, with the proposed strategy being approximately 10^9 times cheaper. These findings (Figure 3) highlight the potential of DNA-based storage as a promising and cost-effective solution for storing large-scale image datasets, particularly in medical imaging.

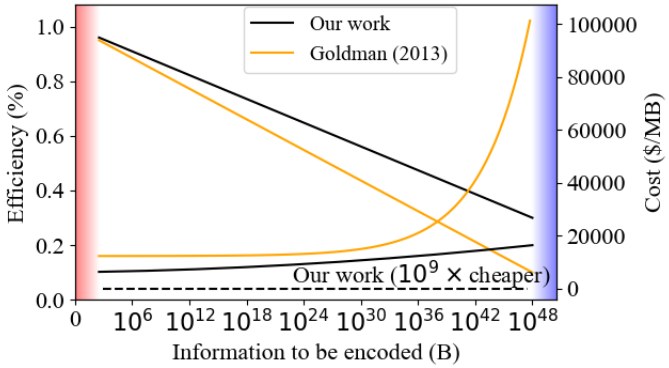

Figure 3. The comparison of the proposed RFS-codec approach with Goldman's work for cost-effectiveness.

5.5 Comparative Analysis

Meanwhile, the proposed approach's performance is compared with prior benchmark studies. Table 5

compares the proposed encoding with an improved density to many other studies. It signifies the RFS-codec's ability to store various larger data files in smaller DNA nucleotides with enhanced densities while satisfying bio-constraints.

6 Conclusion

This work proposes a novel approach (RFS-codec) to combat the challenge of image storage in DNA by avoiding the propagation of nonspecific hybridization errors that sternly affect the durability of stored data. The RFS-codec approach comprises a novel fraction strategy and a new codec method to store image data in DNA. The fraction strategy splits the image to deliver a cost-effective solution, and the codec method provides an encoding to control the errors during the conversion of binary data to DNA bases. The images are encoded and decoded successfully by satisfying two crucial DNA constraints (GC and HP) and attaining 1.8 bit/nt density. The satisfaction of these constraints and high density empowers the proposed approach to control the errors (Figure 2) and cost issues (Figure 3). The results demonstrate significant advantages in constructing cost-effective, highly dense, scalable, and durable DNA data storage.

In the future, the proposed approach will be assessed in wet lab experiments to synthesize and sequence the DNA data for end-to-end DNA data storage systems.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The author declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-generation digital information storage in DNA. *Science, 337*(6102), 1628. [CrossRef]

[2] Erlich, Y., & Zielinski, D. (2017). DNA Fountain enables a robust and efficient storage architecture. *Science, 355*(6328), 950-953. [CrossRef]

[3] Cao, B., Wang, K., Xie, L., Zhang, J., Zhao, Y., Wang, B., & Zheng, P. (2024). PELMI: Realize robust DNA image storage under general errors via parity encoding and local mean iteration. *Briefings in Bioinformatics, 25*(5), bbae463. [CrossRef]

[4] Song, L., Geng, F., Gong, Z. Y., Chen, X., Tang, J., Gong, C., ... & Yuan, Y. J. (2022). Robust data storage in DNA by de Bruijn graph-based de novo strand assembly. *Nature communications, 13*(1), 5361. [CrossRef]

[5] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., & Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *nature, 494*(7435), 77-80. [CrossRef]

[6] Davis, J. (1996). Microvenus. *Art Journal, 55*(1), 70-74. [CrossRef]

[7] Bancroft, C., Bowler, T., Bloom, B., & Clelland, C. T. (2001). Long-term storage of information in DNA. *Science, 293*(5536), 1763-1765. [CrossRef]

[8] Pan, C., Tabatabaei, S. K., Tabatabaei Yazdi, S. M. H., Hernandez, A. G., Schroeder, C. M., & Milenkovic, O. (2022). Rewritable two-dimensional DNA-based data storage with machine learning reconstruction. *Nature Communications, 13*(1), 2984. [CrossRef]

[9] Cao, B., Zheng, Y., Shao, Q., Liu, Z., Xie, L., Zhao, Y., ... & Wei, X. (2024). Efficient data reconstruction: The bottleneck of large-scale application of DNA storage. *Cell Reports, 43*(4). [CrossRef]

[10] Antkowiak, P. L., Lietard, J., Darestani, M. Z., Somoza, M. M., Stark, W. J., Heckel, R., & Grass, R. N. (2020). Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. *Nature communications, 11*(1), 5345. [CrossRef]

[11] Banal, J. L., Shepherd, T. R., Berleant, J., Huang, H., Reyes, M., Ackerman, C. M., ... & Bathe, M. (2021). Random access DNA memory using Boolean search in an archival file storage system. *Nature materials, 20*(9), 1272-1280. [CrossRef]

[12] Zhou, S., Zhang, Q., & Wei, X. (2010). Image encryption algorithm based on DNA sequences for the big image. *2010 International Conference on Multimedia Information Networking and Security*, 884-888. [CrossRef]

[13] Fan, Q., Lilja, D. J., & Sapatnekar, S. S. (2020). Adaptive-length coding of image data for low-cost approximate storage. *IEEE Transactions on Computers, 69*(2), 239-252. [CrossRef]

[14] Li, Q., Shi, L., Yang, J., Zhang, Y., & Xue, C. J. (2019). Leveraging approximate data for robust flash storage. *2019 56th ACM/IEEE Design Automation Conference (DAC)*, 1-6. [CrossRef]

[15] Organick, L., Ang, S. D., Chen, Y. J., Lopez, R., Yekhanin, S., Makarychev, K., ... Strauss, K. (2018). Random access in large-scale DNA data storage. *Nature Biotechnology, 36*(3), 242-248. [CrossRef]

[16] Cao, B., Zhang, X., Cui, S., & Zhang, Q. (2022). Adaptive coding for DNA storage with high storage density and low coverage. *NPJ systems biology and applications*, *8*(1), 23. [CrossRef]

Abdur Rasool is an accomplished researcher in computer science and applied technology, specializing in artificial intelligence, DNA data storage, data security natural language processing, and FinTech applications. He earned his Ph.D. in Computer Applied Technology from the Shenzhen Institute of Advanced Technology (SIAT), University of Chinese Academy of Sciences (UCAS), China, in 2023. Dr. Rasool is currently a Postdoctoral Associate at the University of Hawai'i at Mānoa, USA, where he focuses on developing multimodal machine learning models for diagnosing neurobehavioral conditions using computer vision and human-in-the-loop methodologies. He has authored over 30 peer-reviewed articles in prestigious journals, such as Small Methods.

He has received multiple recognitions for excellence in academia, such as the Shenzhen Universiade International Fund (2022, 2023), the Excellent International Graduate Award (2022, 2023), the IEEE R10 Best Paper Award, and the UROP Funding Program at the University of Hawai'i.

He has contributed significantly to the scientific community as editor-in-chief for the Journal of Artificial Intelligence in Bioinformatics, guest editor for MDPI Electronics, and associate editor for the ICCK Transactions on Advanced Computing and Systems. His extensive peer-review activities for leading journals, including IEEE Transactions on NanoBioscience, Briefings in Bioinformatics, IEEE Journal of Translational Engineering in Health and Medicine, Applied Artificial Intelligence, Journal of Medical Internet Research, The Imaging Science Journal, Frontiers in Genetics, and Journal of Supercomputing. (Email: abdur@hawaii.edu)