frontiers | Frontiers in Bioengineering and Biotechnology

TYPE Original Research
PUBLISHED 19 April 2023
DOI 10.3389/fbioe.2023.1173763

OPEN ACCESS

EDITED BY
Jose Ruben Morones-Ramirez,
Autonomous University of Nuevo León, Mexico

REVIEWED BY
Chanjuan Liu,
Dalian University of Technology, China
Tao Song,
China University of Petroleum, China
Henry Han,
Baylor University, United States

*CORRESPONDENCE
Peng Xu,
gdxupeng@gzhu.edu.cn
Wenbin Liu,
wbliu6910@gzhu.edu.cn

†These authors contributed equally to the work

SPECIALTY SECTION
This article was submitted to Synthetic Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

RECEIVED 25 February 2023
ACCEPTED 30 March 2023
PUBLISHED 19 April 2023

CITATION
Zan X, Chu L, Xie R, Su Y, Yao X, Xu P and Liu W (2023), An image cryptography method by highly error-prone DNA storage channel.
Front. Bioeng. Biotechnol. 11:1173763.
doi: 10.3389/fbioe.2023.1173763

An image cryptography method by highly error-prone DNA storage channel

Xiangzhen Zan[1†], Ling Chu[1†], Ranze Xie[1], Yanqing Su[1], Xiangyu Yao[1], Peng Xu[1,2,3*] and Wenbin Liu[1,3*]

[1]Institute of Computational Science and Technology, Guangzhou University, Guangzhou, Guangdong, China, [2]School of Computer Science of Information Technology, Qiannan Normal University for Nationalities, Duyun, Guizhou, China, [3]Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou, Guangdong, China

Introduction: Rapid development in synthetic technologies has boosted DNA as a potential medium for large-scale data storage. Meanwhile, how to implement data security in the DNA storage system is still an unsolved problem.

Methods: In this article, we propose an image encryption method based on the modulation-based storage architecture. The key idea is to take advantage of the unpredictable modulation signals to encrypt images in highly error-prone DNA storage channels.

Results and Discussion: Numerical results have demonstrated that our image encryption method is feasible and effective with excellent security against various attacks (statistical, differential, noise, and data loss). When compared with other methods such as the hybridization reactions of DNA molecules, the proposed method is more reliable and feasible for large-scale applications.

KEYWORDS

image encryption, DNA storage, highly error-prone DNA storage channel, multiple sequence alignment, information security

1 Introduction

As the storage medium of genetic information, DNA molecules have the advantage of long durability, high density, and low cost. Recent advancements in their synthesis and sequencing technologies have made DNA a promising medium to deal with the challenges of data explosion (Qian et al., 2020; Meiser et al., 2022). Currently, researchers have devoted a lot of effort to accurately recover information from the noised sequence pool (Erlich and Zielinski, 2017; Meiser et al., 2019; Antkowiak et al., 2020; Press et al., 2020; Jeong et al., 2021). However, how to ensure the security of private data in DNA storage is an important question that is still in its infancy.

Clelland et al. (1999) first hid some secret letters in microdots of DNA molecules. Later, Gehani et al. (2004) realized the one-time pad encryption on DNA molecules through DNA microarray technology. In the past decade, researchers have continued to explore the encryption potential of complex biochemical processes. Yang et al. (2014) implemented a 32-bit one-time pad encryption that simulated one-bit exclusive-OR (XOR) operation by DNA strand displacement reaction (SDR). Later, Peng et al. (2018) developed a three-dimensional DNA self-assembly pyramid structure to achieve double-bit encryption. Zhang et al. (2019) constructed a DNA origami cryptography method by folding M13 viral scaffolds which could communicate braille-like patterns at the nanometer scale. Zakeri et al. (2016)

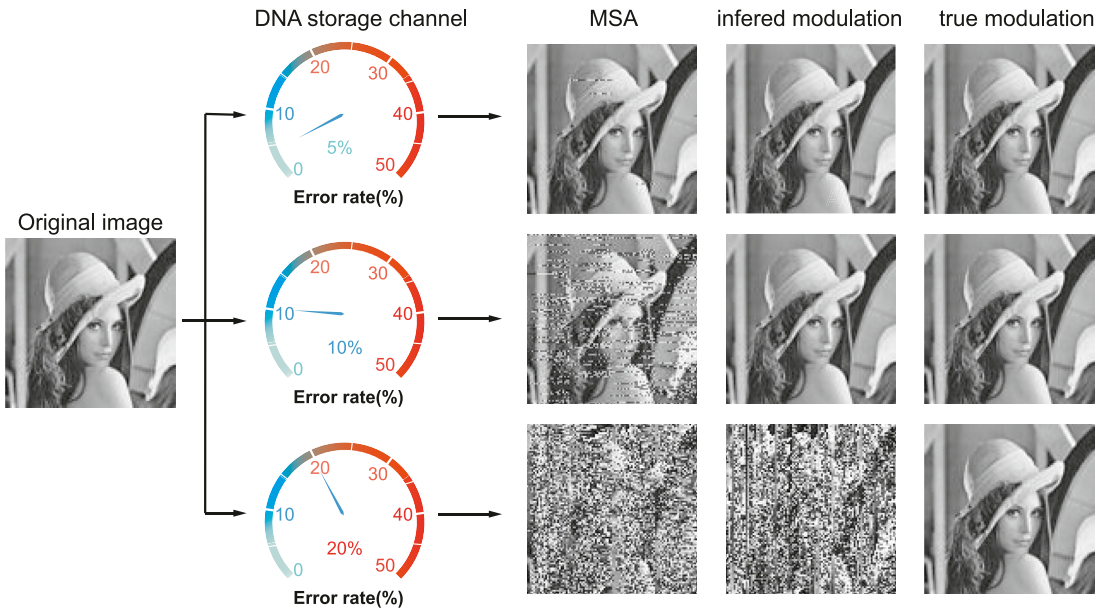

FIGURE 1
Recovered images at error rate 5%, 10%, and 20% by MSA, inferred, and true modulation signal. The recovered images by the first two methods gradually become vague as the error rate increases, while the one by the true modulation signal is completely correct.

accomplished short message communication by chromatogram patterning and multiplexed DNA sequence encoding technology. Peng et al. (2021) proposed a one-time-pad cipher algorithm by confusion mapping and random adapter, which could guarantee controllable biological security. Recently, some researchers also developed an SDR-based chaos system to generate secret keys (Wang et al., 2020; Zou et al., 2021; Zhu et al., 2022). However, the reliability and practicability of these methods are limited in two aspects. First, they are vulnerable to the base errors that are prevalent in DNA storage. Due to the over-reliance on highly specific biomolecule hybridization reactions, these methods require specialized design and accurate synthesis of DNA sequences. Even a few base errors can cause encryption failure. Second, the experiments are sophisticated and may produce unpredictable results in case of some subtle variations in experiment conditions (temperature, time, and ion concentration). Moreover, noise environments may even worsen the unpredictability of the results. In addition, these experimental processes are time-consuming, difficult to monitor, and not suitable for large-scale applications.

Recently, our group proposed a modulation-based DNA storage architecture that is extremely robust to insertion–deletion–substitution (IDS) errors. The basic idea is that the modulation signal not only converts the binary information into DNA sequences during the write phase but also detects synchronization errors and decodes the corrected data during the read phase (Zan et al., 2022). Figure 1 shows an example of the recovered image under different noise levels by three strategies. The first one reconstructs the images directly by multiple sequence alignment (MSA) algorithms. The second one infers a possible modulation signal M' by MSA and then reconstructs the images using the inferred M' as in Zan et al. (2022). While the last one recovers images using the true modulation signal M. As noises increase, the first two gradually fail to recover the original image while the last one could perfectly recover it. Since MSA is the only method available for noise correction without the knowledge of coding in DNA storage and since IDS errors are inherent in the synthesis and sequencing processes, the modulation signal could serve as a secure key in a high-error DNA storage channel.

In this article, we explore the feasibility of image encryption in a high DNA storage channel. The proposed image encryption scheme consists of two layers: conventional encryption and DNA storage channel encryption. The first layer implements pixel scrambling and diffusion, and the second layer adds further complex confusion to DNA sequences (or DNA pixels) by taking advantage of the uncertainty in the DNA storage channel. Simulation results have demonstrated that the proposed method could resist cipher attack at the DNA sequence level when the noise is larger than 20%. It is also very robust to DNA base errors and sequence losses. Security analysis proves that it has a large key space, is sensitive to the key and plaintext, and can cope with statistical attacks. In sum, the proposed method achieves an excellent combination of the silico-based and carbon-based information security technologies and paves a solid foundation for data security in future DNA-based information architecture.

2 Encryption and decryption

Figure 2 shows the schematic diagram of the proposed encryption and decryption processes, which includes two stages. The first stage performs regular pixel scrambling and diffusion at the binary level. The second stage further encrypts the binary data into DNA sequences by a known modulation key; then, these are transmitted through the highly error-prone DNA storage

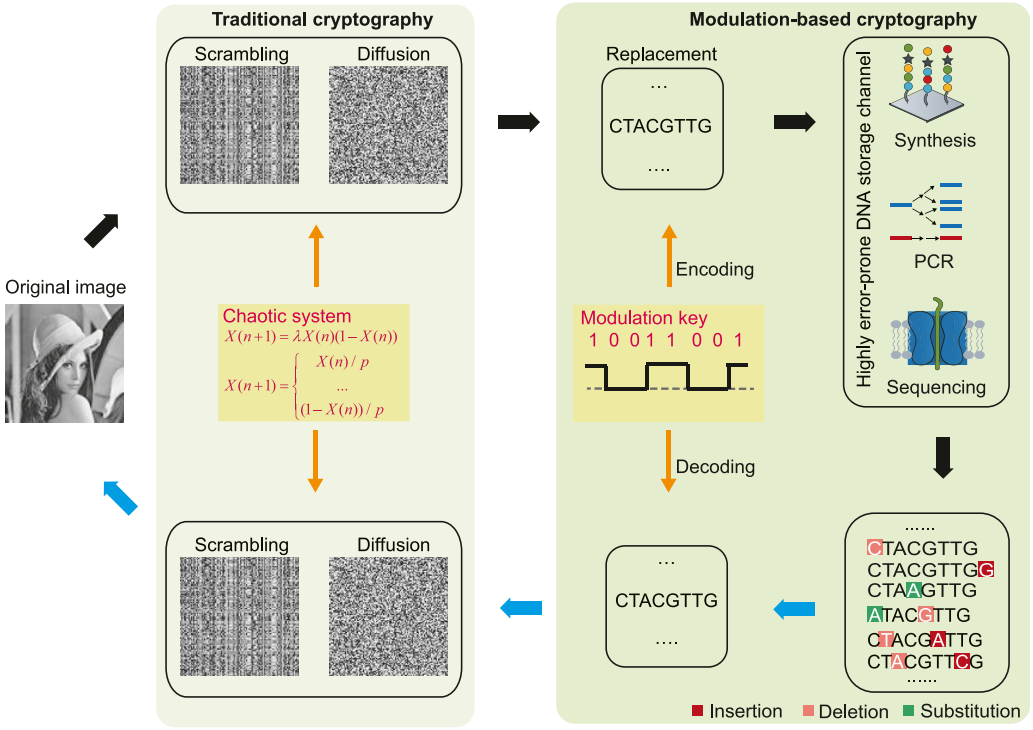

FIGURE 2
Schematic diagram of encryption and decryption processes. To the left is traditional cryptography which includes scrambling and diffusion at the binary level. To the right is modulation-based cryptography at the highly error-prone DNA storage channel. Black arrows represent the encryption process, while blue arrows indicate the decryption process.

channel, which consists of several DNA operating technologies with high error rates, such as light-directed maskless array DNA synthesis with an error rate of approximately 15% (Antkowiak et al., 2020), biased polymerase chain reaction (PCR) (Chen et al., 2020), and nanopore sequencing with error rates between 10% and 15% (Wang et al., 2021). Finally, the output ciphertext is a pool of DNA sequences involving large amounts of insertion–deletion–substitution errors. The decryption process is the reverse of encryption.

2.1 Secret key generation

Secret keys mainly consist of two parts. One is the chaotic systems which include the piecewise linear chaotic map (PWLCM) (Alawida et al., 2019; Zhou et al., 2021) and logistic map (Sui et al., 2015; 2014), while the other is the modulation key.

The dynamic equation of PWLCM can be described by the following function:

$$X(n+1)=\begin{cases} X(n)/p, 0\le X(n)<p \\ (X(n)-p)/(0.5-p), p\le X(n)<0.5 \\ (1-X(n)-p)/(0.5-p), 0.5\le X(n)<1-p \\ (1-X(n))/p, 1-p\le X(n)\le 1 \end{cases} \quad (1)$$

where the parameter p should be in the range of (0, 0.5), and the status value $X(n)$ is in the range of (0,1).

The logistic map is defined as follows:

$$X(n+1)=\lambda X(n)(1-X(n)) \quad (2)$$

where the parameter λ should be in the range of (0, 4), and the status value $X(n)$ is in the range of (0,1).

We use the abovementioned chaotic systems to generate three random sequences, two of which are generated by the PWLCM with the initial status values $X_r(0)$ and $X_c(0)$ and one by a logistic map with the initial status value $X_d(0)$. To relate the initial values with the plain image, we use Keccak (Bertoni et al., 2013) to hash the plain image to generate a fixed-length K (512 bit), which can be divided into 32 blocks, each of 16-bit. We denote it as $K=\{k_1, k_2, \ldots, k_{32}\}$. The initial status values are derived as follows:

$$\begin{cases} X_r(0)=\frac{1}{16}\sum_{i=k_1}^{k_{11}}\left(\frac{32}{4k_i+1}+\frac{1}{4k_i+3}+\frac{128}{10*k_i+5}+\frac{64}{10k_i+7}+\frac{4}{10k_i+9}+\frac{1}{10k_i+3}\right) \\ X_c(0)=\frac{1}{16}\sum_{i=k_{12}}^{k_{22}}\left(\frac{32}{4k_i+1}+\frac{1}{4k_i+3}+\frac{128}{10*k_i+5}+\frac{64}{10k_i+7}+\frac{4}{10k_i+9}+\frac{1}{10k_i+3}\right) \\ X_d(0)=\frac{1}{16}\sum_{i=k_{23}}^{k_{32}}\left(\frac{32}{4k_i+1}+\frac{1}{4k_i+3}+\frac{128}{10*k_i+5}+\frac{64}{10k_i+7}+\frac{4}{10k_i+9}+\frac{1}{10k_i+3}\right) \end{cases} \quad (3)$$

After retrieving the initial value [i.e., X_r (0)] and the corresponding chaotic map, we iterate through the chaotic map n times [(i.e., $X_r(n)$] to remove transient processes and then continue to iterate it to obtain the random sequence of the specified length.

Modulation key M is a binary sequence of equal length to the encoded DNA sequence. In M, '0' represents A/T and '1' represents C/G (Zan et al., 2022). The 01 composition of the key directly reflects the base

composition of the encoded DNA sequence. Since the DNA sequences with extreme guanine–cytosine (GC) content or long homopolymers (i.e., longer repeats of the same base, i.e., AAAAAA. . .) are difficult to synthesize and prone to sequencing errors, most of the DNA storage works comply with some encoding constraints on the DNA sequences, such as the GC content of 45%–55% and homopolymer runs of ≤ 3nt (Erlich and Zielinski, 2017). In our encryption scheme, the percentage of 1 s (or 0 s) in the modulation key (equivalent to the GC content) and the consecutive length of 1 s (or 0 s) (equivalent to the homopolymer runs) also adhere to these constraints. From a key space perspective, this makes key cracking more difficult.

2.2 Encryption algorithm

Given an image P with size $W \times H$ and the iteration number $n \in [100, \infty)$. Let $N = W \times H$, the detailed encryption process can be depicted as follows.

2.2.1 Traditional cryptography by scrambling and diffusion

Step 1. Get the secret keys λ, p, $X_r(0)$, $X_c(0)$, and $X_d(0)$.

Step 2. Use $X_r(0)$, n, and Eq. 1 to generate one sequence S_R of length W. Sort S_R in the ascending order to get the corresponding index sequence S_R', number the rows of pixels of the original image P, and adjust row positions according to S_R' to finish row-wise permutation operations. The row-wise scrambled image is denoted as P_1. For example, let $S_R = \{35, 60, 13\}$ and image $P = \{r_1, r_2, r_3\}$, where r_i $(1 \le i \le 3)$ stands for the *i-th* row of pixels, the corresponding index sequence is $S_R' = \{3, 1, 2\}$, and the row-wise scrambled image is $P_1 = \{r_3, r_2, r_1\}$.

Step 3. Similar to Step 2, use $X_c(0)$, n, and Eq. 1 to generate one sequence S_C of length H and perform column-wise permutation operations on P_1. The scrambled image is denoted as P_2.

Step 4. Use $X_d(0)$, n, and Eq. 2 to generate a sequence D of length of $W \times H$. Reshape P_2 into one-dimensional sequence Q. Performing diffusion operation on Q using Eq. 4 yields Q'. Finally, reshape Q' into a two-dimensional $W \times H$ matrix P_3.

$$Q'(i) = \begin{cases} (Q(1) \oplus Q(N) \oplus Q(N-1) \oplus D(1)) mod 256, & i = 1 \\ (Q(2) \oplus Q'(1) \oplus Q(1) \oplus D(1)) mod 256, & i = 2 \\ (Q(i) \oplus Q'(i-1) \oplus Q'(i-2) \oplus D(i)) mod 256, & i \in [3, N] \end{cases} \tag{4}$$

2.2.2 Dynamic modulation cryptography

Step 1. Obtain the secret key M.

Step 2. Transform P_3 into the binary form P_3', and partition P_3' into strands of fixed length ($l = len(M)$). All these strands are encrypted with M to generate their corresponding DNA sequences C according to a simple mapping rule (00 → A, 01 → T, 10 → C, 11 → G). For instance, assuming M = '**1001**1001**1001**', the message strand '010011010110' is aligned with M into two rows, and a DNA sequence '**CTAC**GTAG**CTTC**' can be obtained after mapping each column of the two rows into one DNA base.

Step 3. Transform C into the final ciphertext C' through the highly error-prone DNA storage channel.

2.3 Decryption algorithm

As an asymmetric cryptosystem is more secure than a symmetric cryptosystem (Dong, 2015), the decryption keys are not identical to the encryption ones in our method. The decryption scheme uses the keys λ, p, $X_r(n)$, $X_c(n)$, $X_d(n)$, and M to execute the reverse operation on the encryption algorithm. First, according to the modulation decoding method (Zan et al., 2022), M is used to correct noises in the sequenced data C' and decode them to obtain the two-dimensional pixel matrix P_3. Second, Eqs. 4, 2, λ, and $X_d(n)$ are used to perform reverse diffusion operations on P_3 to get P_2. Finally, Eq. 1, p, $X_r(n)$, and $X_c(n)$ are used to perform reverse scrambling operations on P_2 to derive plain image P.

3 Results

We demonstrate our results on the 100 × 100 Lena image as a proof of concept. It is encoded by 400 DNA sequences of 200 bases without considering overheads of the index because we assume that the clustering accuracy can be perfect. To investigate the proper noise channel for robust encryption, we take a series of simulation experiments with noises ranging from 2% to 40% and sequence copies ranging from 5 to 10,000.

3.1 Key space analysis

The key space of the proposed method is sufficiently large to withstand any brute force attack. In the traditional decryption process, the receiver has to know the five parameters λ, p, $X_r(n)$, $X_c(n)$, and $X_d(n)$. As their valid precision is 10^{-16}, the key space of the five parameters will be

$$S_{key} = 10^{80} \approx 2^{266} \tag{5}$$

Given that the sequence length is 200, and the percentage of 1s in the carrier strand is about 0.5, the modulation key space is

$$S_M = \binom{200}{100} \approx 2^{196} \tag{6}$$

The total key space of our method is

$$S = S_{key} \times S_M = 10^{80} \times \binom{200}{100} \approx 2^{462} \gg 2^{128} \tag{7}$$

It is much larger than the theoretical secure key value 2^{128} (Dong et al., 2022). As the modulation key space alone is larger than 2^{128}, we can conclude that the storage channel can serve as another layer for data security.

3.2 Ciphertext attack in DNA sequence level

Attackers have two possible ways to decipher the encrypted image in the noisy DNA storage channel. One is to infer a possible

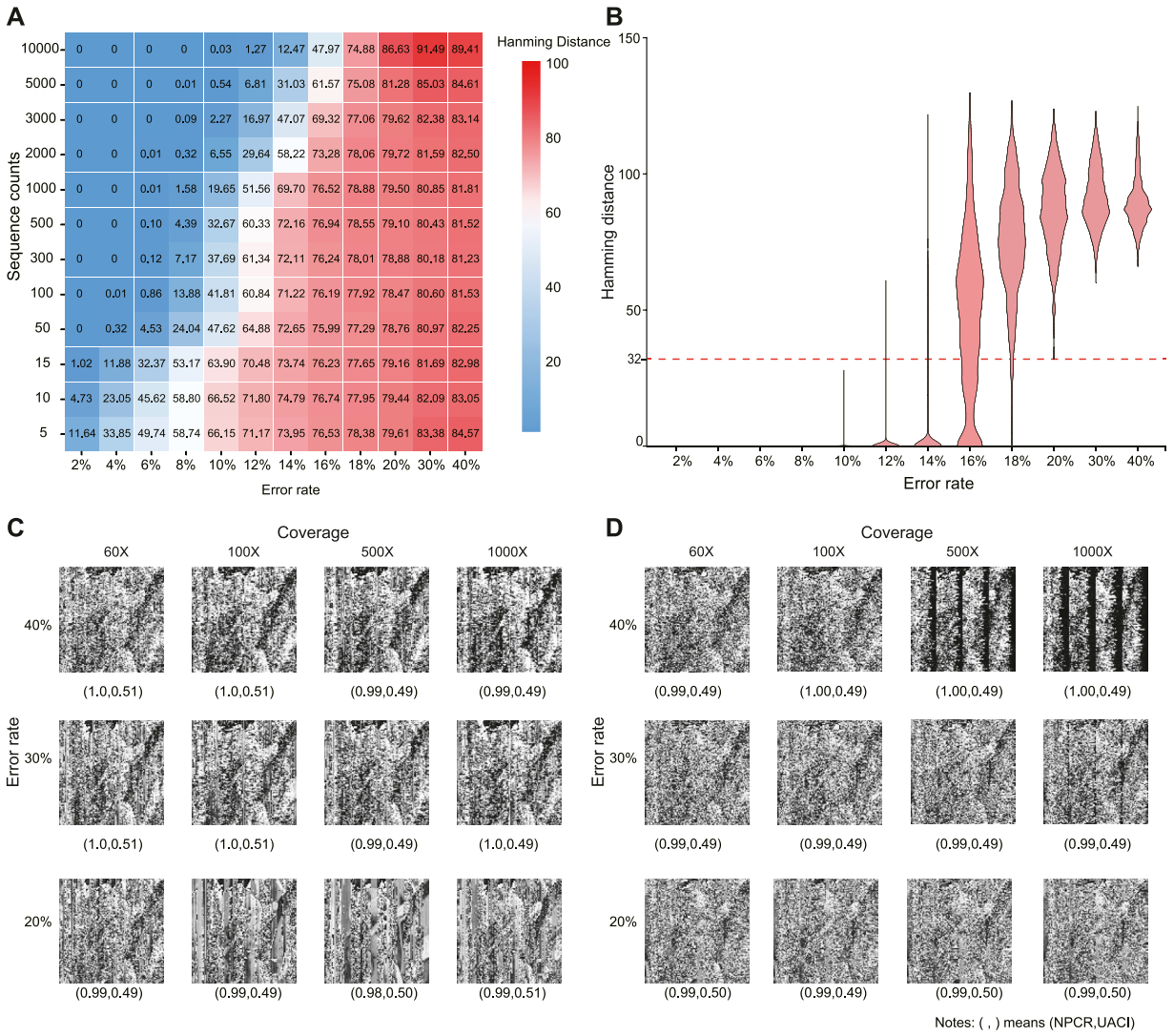

FIGURE 3
Cipher attack at different error rates and sequence copies. **(A)** Average Hamming distance between inferred and true modulation keys. **(B)** Distribution of Hamming distances of inferred modulation keys at sequence copies 10,000. **(C)** Decrypted images by inferred modulation keys. **(D)** Decrypted images by MSA. In **(C)** and **(D)**, the values in parentheses denote NPCR and UACI, respectively.

modulation key M' by MSA and then decipher the sequenced reads by it, and the other is to directly decipher sequenced reads by the MSA algorithm. This is because there are only two methods of correcting base errors in DNA storage: constraint coding and multiple sequence alignment (MSA) without prior knowledge. As MSA fundamentally relies on pairwise sequence alignment algorithms, such as the Needleman–Wunsch algorithm (Needleman, 1970) and seeks to find a globally optimal alignment between multiple copies, there is limited variability in alignment accuracy across MSA software tools (Pervez et al., 2014). Assuming all keys are known except for M, we apply one of the famous MSA tools named MAFFT (Katoh et al., 2002) to conduct a series of experiments.

It is impossible to infer a potential modulation key when the error rate is higher than 20%. The attacker can decipher the encrypted image if the inferred key M' is very similar to M. Here, we assume that the attackers could have sufficient sequence copies to infer M. Figure 3A shows that the average Hamming distance between M and M' increases as the error rate increases. When the error rate is larger than 20%, the average Hamming distance is about 80, and increasing sequence copies may even result in a larger Hamming distance (see the top left corner). Figure 3B further shows the Hamming distance distribution at 10,000 sequence copies. The least Hamming distance may reach 32 at an error rate of 20%. That is, there are at least 32 errors in the inferred modulation keys with 200 bits. As the error rate increases, this lower limit could further increase. Therefore, inferring the true modulation key becomes almost impossible in a high error channel.

Without knowing the modulation key M, it is almost impossible to decipher the real image when the error rate is larger than 20%. To evaluate the difference between the decrypted and original images, the number of pixels change rate (NPCR) and unified average changing intensity (UACI) are calculated as

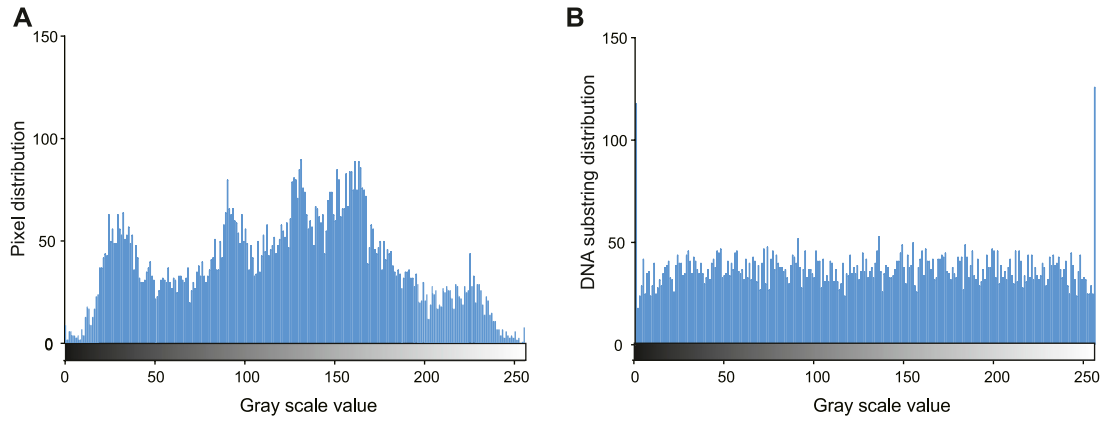

FIGURE 4
Distribution of pixel intensity histogram. **(A)** Plain image. **(B)** Encrypted image.

$$D(i,j) = \begin{cases} 1, c_1(i,j) \neq c_2(i,j) \\ 0, otherwise \end{cases} \quad (8)$$

$$NPCR = \frac{\sum_{ij} D(i,j)}{W \times H} \times 100 \quad (9)$$

$$UACI = \frac{1}{W \times H}\left[\sum_{ij} \frac{|c_1(i,j) - c_2(i,j)|}{255} \times 100\%\right] \quad (10)$$

where W and H are the width and height of two images (c_1 and c_2), respectively. Figures 3C, D show the decrypted images using different sequence copies by the inferred modulation key and MSA, respectively. Compared with the original image, the decrypted images are all seriously distorted with $NPCR \approx 1$ and $UACI \approx 0.5$, even at sequence copies 1,000. The utilization of the traditional cryptographic techniques further increases crack difficulties.

3.3 Sensitivity analysis

The proposed method is sensitive to secrete keys and plaintext. A slight change in the key (i.e., a single bit change) or plaintext could cause a completely different encrypted result. First, the sensitivity of the PWLCM and logistic map has been confirmed in many image-encryption works (Sui et al., 2015; Sui et al., 2014; Alawida et al., 2019; Zhou et al., 2021). At the same time, 1 bit insertion/deletion in the modulation signal will affect the encoding of a large number of pixels. Second, plaintext sensitivity is accomplished by the pixel diffusion process and initial status values of the chaotic systems which are strongly related to the plain image.

3.4 Statistical analysis

The proposed method can resist statistical attacks. Figure 4 shows the histogram of the pixels in the original image (A) and the encoded eight-base pixel DNA strands (B). The distribution of the encoded DNA sequences is more flat than that of the original. Considering the IDS noises in the sequenced reads, the distribution in (B) tends to be more uniform. Table 1 shows the correlation coefficients of the ciphered image after dislocation and diffusion. All values in the three directions are close to the ideal value of 0 (Ghadirli et al., 2019). That is, the encrypted pixels are distributed randomly. The information entropy of the cyphered image is 7.950121813, which is very close to the ideal value of 8 (Ghadirli et al., 2019). Therefore, the encrypted image shows favorable randomness.

TABLE 1 Correlation coefficients in different directions of original and ciphered images.

Image	Horizontal	Vertical	Diagonal
Original	0.873734246	0.945931872	0.827460129
Ciphered	−0.011323434	−0.010079104	0.007942569

3.5 Robustness analysis

The proposed method is robust to the two most commonly seen errors in DNA storage: base errors and sequence loss. Sequence loss refers to the loss of some DNA molecules during DNA storage processes (e.g., DNA decay, PCR, and sequencing) due to the complexity of the biochemical reactions. Figure 5A shows the decrypted images at an error rate of 20~ 40% and sequence copies 50~1,000. The original images could be completely deciphered, given sufficient sequence copies. Figure 5B shows the decrypted images which could retain the portrait even at a loss rate of 50%. It should be added that the proposed method can easily be combined with an erasure code, such as a fountain code (Mackay, 2005), to further improve its resistance to sequence loss attacks. The combined method is quite simple. All that is required is to encode P_3 with a fountain code prior to dynamic modulation encryption. To the best of our knowledge, such robustness can only be achieved by modulation-based DNA storage architecture (Yazdi et al., 2017; Antkowiak et al., 2020; Press et al., 2020; Srinivasavaradhan et al., 2021; Song et al., 2022; Zan et al., 2022).

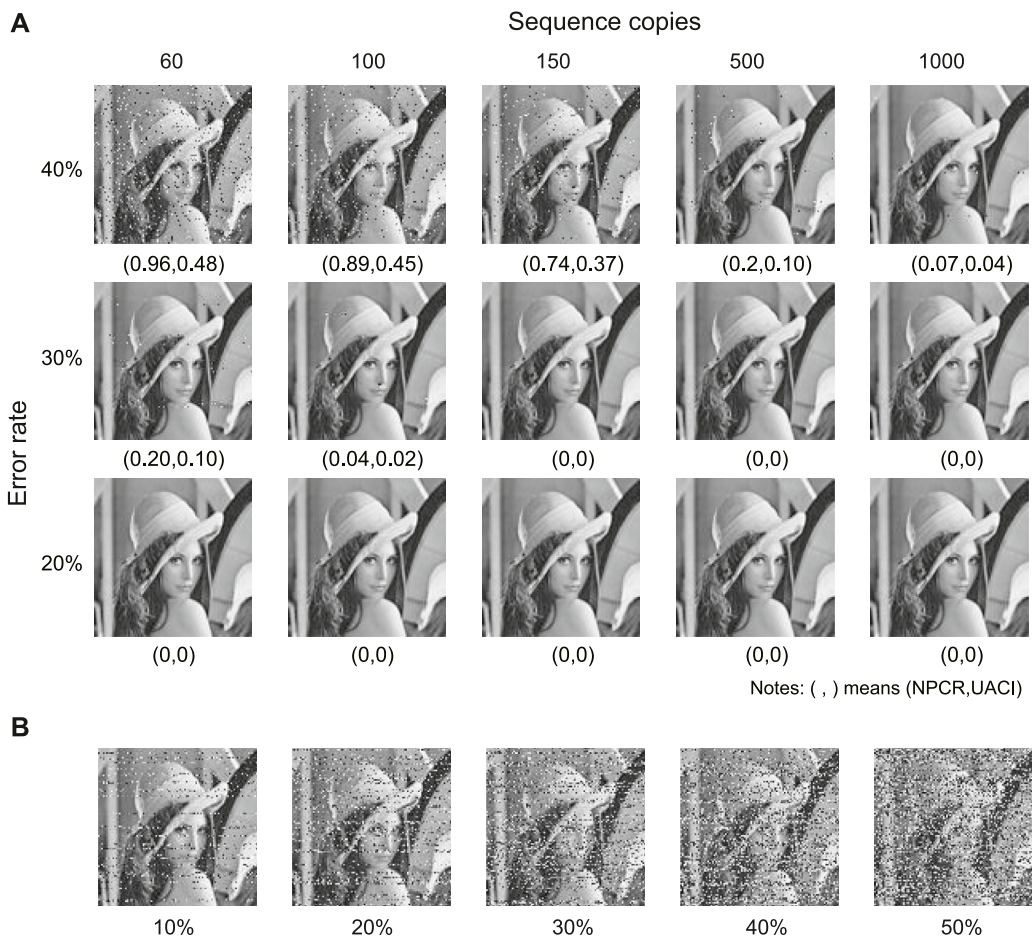

FIGURE 5
Robustness to base errors and sequence loss. **(A)** Decrypted images at different error rates and sequence copies. **(B)** Decrypted images at different sequence loss rates.

TABLE 2 Comparisons of encryption methods for DNA storage.

Literatures	Dynamic encoding	Dynamic encryption	Robustness	Biological encryption	Large-scale encryption	Logical density (bits/nt)	Key space
Yang et al. (2014)	√	*	*	√	×	0.006	$\binom{4^{25}\times 1000}{2000}$
Zakeri et al. (2016)	×	*	*	√	×	0.239	9.1×10^{61}
Zhang et al. (2019)	×	*	*	√	×	0.001	2^{702}
Peng et al. (2021)	√	√	×	√	×	1.65	2^{400}
Zhu et al. (2022)	√	√	×	√	×	2.0	$2^{1,536}$
This work	√	√	√	√	√	1.0	2^{462}

×, indication of minimum level of support; √, indication of acceptable level of support; *, partial fulfillment.

3.6 Classical attack analysis

The proposed method is resistant to classical attacks, such as known plaintext attacks, chosen plaintext attacks, and chosen ciphertext attacks. As mentioned earlier, the secret keys depend not only on the given initial values, such as modulation keys and system parameters, but also on the plain image. For every plain image, the keys are changed both in the encryption process and

decryption process. As such, attackers cannot extract any useful information, either by encrypting a pre-designed special image or by decrypting a certain ciphertext. This concludes that chosen plaintext, chosen ciphertext, and known plaintext attacks do not work against the proposed method.

3.7 Comparisons with other methods

Table 2 shows the detailed comparisons of existing studies. When compared with other methods, our method has the following advantages in terms of encryption using DNA molecules: first, the modulation key and chaotic systems feature our encryption scheme with dynamic encoding and encryption, which can withstand any kind of brute force attack. More importantly, modulation encoding provides a natural way to comply with biochemical constraints for long-term storage. Second, encrypting data by noise storage channels avoids the complexity and uncertainty in biochemical reactions, such as DNA strand displacement, DNA hiding, and DNA self-assembly. Finally, it is the only method with both high logical information density and strong robustness, which can tolerate extreme environments with high base noise and sequence loss. We believe that all these features endow our method with the potential to achieve reliable, secure, robust, and scalable encryption for DNA storage.

4 Conclusion

We propose an image encryption method for DNA storage which includes two parts: conventional encryption and DNA storage channel encryption. The proposed method highlights the importance of unpredicted modulation signals in a highly error-prone DNA storage channel. Simulation results show that our method is feasible and effective for encrypting and decrypting images when the error rate of the DNA storage channel is higher than 20%. There are two ways to generate such high noise: one is to adopt high-error DNA operating technologies, such as light-directed maskless array DNA synthesis, biased PCR, and nanopore sequencing; the other is to construct multiple substitution-rich copies prior to DNA synthesis with an error rate of 20% for each coding sequence. Further analysis of the security shows that it is sensitive to both keys and plaintexts, has a large enough key space, and can resist various attacks (i.e., statistical, only ciphertext, noise and data loss, etc.). When compared with other state-of-the-art encryption methods, our approach has high logical information density, compliance with biochemical constraints, and strong robustness to base errors and sequence loss; it is thus more suitable for large-scale DNA encryption storage. Although designed for image encryption, our method can also be applied to other areas of encryption. Relying on the powerful error correction capability of the modulation-based DNA storage architecture, we believe our approach will further accelerate the arrival of large-scale DNA encrypted storage.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material; further inquiries can be directed to the corresponding author.

Author contributions

XZ and WL conceived the concept. XZ wrote the Python codes. XZ and XY conducted simulation and data analysis. XZ and YS prepared the figures and tables. XZ and LC drafted the manuscript. RX and PX revised the manuscript. WL and PX supervised the study. All authors read and approved the final manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China (No. 62072128 and 24562002079), the Natural Science Foundation of Guangdong Province of China (No. 2023A1515011401), and the Open Project of Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application (No. 2022B1212010011).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, editors, and reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alawida, M., Samsudin, A., Teh, J. S., and Alkhawaldeh, R. S. (2019). A new hybrid digital chaotic system with applications in image encryption. *Signal Process.* 160, 45–58. doi:10.1016/j.sigpro.2019.02.016

Antkowiak, P. L., Lietard, J., Darestani, M. Z., Somoza, M. M., Stark, W. J., Heckel, R., et al. (2020). Low cost dna data storage using photolithographic synthesis and advanced information reconstruction and error correction. *Nat. Commun.* 11, 5345. doi:10.1038/s41467-020-19148-3

Bertoni, G., Daemen, J., Peeters, M., and Van Assche, G. (2013). *Keccak. Times cited in web of science core collection: 85 total times cited: 90 cited reference count: 11.*

Chen, Y.-J., Takahashi, C. N., Organick, L., Bee, C., Ang, S. D., Weiss, P., et al. (2020). Quantifying molecular bias in dna data storage. *Nat. Commun.* 11, 3264. doi:10.1038/s41467-020-16958-3

Clelland, C. T., Risca, V., and Bancroft, C. (1999). Hiding messages in dna microdots. *Nature* 402, 533–534. doi:10.1038/21092

Dong, C. (2015). Asymmetric color image encryption scheme using discrete-time map and hash value. *Optik* 126, 2571–2575. doi:10.1016/j.ijleo.2015.06.035

Dong, Y., Zhao, G., Ma, Y., Pan, Z., and Wu, R. (2022). A novel image encryption scheme based on pseudo-random coupled map lattices with hybrid elementary cellular automata. *Inf. Sci.* 593, 121–154. doi:10.1016/j.ins.2022.01.031

Erlich, Y., and Zielinski, D. (2017). Dna fountain enables a robust and efficient storage architecture. *Science* 355, 950–954. doi:10.1126/science.aaj2038

Gehani, A., LaBean, T., and Reif, J. (2004). *DNA-Based cryptography*. Berlin, Heidelberg: Springer Berlin Heidelberg, 167–188. doi:10.1007/978-3-540-24635-0_12

Ghadirli, H. M., Nodehi, A., and Enayatifar, R. (2019). An overview of encryption algorithms in color images. *Signal Process.* 164, 163–185. doi:10.1016/j.sigpro.2019.06.010

Jeong, J., Park, S. J., Kim, J. W., No, J. S., Jeon, H. H., Lee, J. W., et al. (2021). Cooperative sequence clustering and decoding for dna storage system with fountain codes. *Bioinformatics* 37, 3136–3143. doi:10.1093/bioinformatics/btab246

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). Mafft: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi:10.1093/nar/gkf436

Mackay, D. (2005). Fountain codes. *IEE Proc. Part I, Commun.* 152, 1062–1068. doi:10.1049/ip-com:20050237

Meiser, L. C., Antkowiak, P. L., Koch, J., Chen, W. D., Kohll, A. X., Stark, W. J., et al. (2019). Reading and writing digital data in dna. *Nat. Protoc.* 15, 86–101. doi:10.1038/s41596-019-0244-5

Meiser, L. C., Nguyen, B. H., Chen, Y.-J., Nivala, J., Strauss, K., Ceze, L., et al. (2022). Synthetic dna applications in information technology. *Nat. Commun.* 13, 352. doi:10.1038/s41467-021-27846-9

Needleman, S., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453. doi:10.1016/0022-2836(70)90057-4

Peng, W., Cheng, D., and Song, C. (2018). One-time-pad cryptography scheme based on a three-dimensional dna self-assembly pyramid structure. *PLOS ONE* 13, e0206612. doi:10.1371/journal.pone.0206612

Peng, W., Cui, S., and Song, C. (2021). One-time-pad cipher algorithm based on confusion mapping and dna storage technology. *PLoS ONE* 16, e0245506. doi:10.1371/journal.pone.0245506

Pervez, M. T., Babar, M. E., Nadeem, A., Aslam, M., Awan, A. R., Aslam, N., et al. (2014). Evaluating the accuracy and efficiency of multiple sequence alignment methods. *Evol. Bioinforma.* 10, EBO.S19199. doi:10.4137/EBO.S19199

Press, W. H., Hawkins, J. A., Jones, S. K., Schaub, J. M., and Finkelstein, I. J. (2020). Hedges error-correcting code for dna storage corrects indels and allows sequence constraints. *Proc. Natl. Acad. Sci. U. S. A.* 117, 18489–18496. doi:10.1073/pnas.2004821117

Qian, L., Ouyang, Q., Ping, Z., Sun, F., and Dong, Y. (2020). Dna storage: Research landscape and future prospects. *Natl. Sci. Rev.* 7, 1092–1107. doi:10.1093/nsr/nwaa007

Song, L., Geng, F., Gong, Z.-Y., Chen, X., Tang, J., Gong, C., et al. (2022). Robust data storage in dna by de bruijn graph-based de novo strand assembly. *Nat. Commun.* 13, 5361. doi:10.1038/s41467-022-33046-w

Srinivasavaradhan, S. R., Gopi, S., Pfister, H., and Yekhanin, S. (2021). *Trellis bma: Coded trace reconstruction on ids channels for dna storage*.

Sui, L., Duan, K., Liang, J., Zhang, Z., and Meng, H. (2014). Asymmetric multiple-image encryption based on coupled logistic maps in fractional Fourier transform domain. *Opt. Lasers Eng.* 62, 139–152. doi:10.1016/j.optlaseng.2014.06.003

Sui, L., Liu, B., Wang, Q., Li, Y., and Liang, J. (2015). Color image encryption by using yang-gu mixture amplitude-phase retrieval algorithm in gyrator transform domain and two-dimensional sine logistic modulation map. *Opt. Lasers Eng.* 75, 17–26. doi:10.1016/j.optlaseng.2015.06.005

Wang, Y., Li, Z., and Sun, J. (2020). Three-variable chaotic oscillatory system based on dna strand displacement and its coupling combination synchronization. *IEEE Trans. NanoBioscience* 19, 434–445. doi:10.1109/TNB.2020.2989577

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 39, 1348–1365. doi:10.1038/s41587-021-01108-x

Yang, J., Ma, J., Liu, S., and Zhang, C. (2014). A molecular cryptography model based on structures of dna self-assembly. *Chin. Sci. Bull.* 59, 1192–1198. doi:10.1007/s11434-014-0170-4

Yazdi, S. M. H. T., Gabrys, R., and Milenkovic, O. (2017). Portable and error-free dna-based data storage. *Sci. Rep.* 7, 5011. doi:10.1038/s41598-017-05188-1

Zakeri, B., Carr, P. A., and Lu, T. K. (2016). Multiplexed sequence encoding: A framework for dna communication. *PLoS One* 11, e0152774. doi:10.1371/journal.pone.0152774

Zan, X., Xie, R., Yao, X., Xu, P., and Liu, W. (2022). A robust and efficient dna storage architecture based on modulation encoding and decoding. *bioRxiv* 2022, 490755. doi:10.1101/2022.05.25.490755

Zhang, Y., Wang, F., Chao, J., Xie, M., Liu, H., Pan, M., et al. (2019). Dna origami cryptography for secure communication. *Nat. Commun.* 10, 5469. doi:10.1038/s41467-019-13517-3

Zhou, P., Du, J., Zhou, K., and Wei, S. (2021). 2d mixed pseudo-random coupling ps map lattice and its application in s-box generation. *Nonlinear Dyn.* 103, 1151–1166. doi:10.1007/s11071-020-06098-0

Zhu, E., Luo, X., Liu, C., and Chen, C. (2022). An operational dna strand displacement encryption approach. *Nanomater. (Basel)* 12, 877. doi:10.3390/nano12050877

Zou, C., Wei, X., Zhang, Q., Zhou, C., and Zhou, S. (2021). Encryption algorithm based on dna strand displacement and dna sequence operation. *IEEE Trans. NanoBioscience* 20, 223–234. doi:10.1109/TNB.2021.3058399