BMC Genomics

RESEARCH

Open Access

DNA-QLC: an efficient and reliable image encoding scheme for DNA storage

Yanfen Zheng[1], Ben Cao[1], Xiaokang Zhang[1], Shuang Cui[1], Bin Wang[2] and Qiang Zhang[1*]

Abstract

Background DNA storage has the advantages of large capacity, long-term stability, and low power consumption relative to other storage mediums, making it a promising new storage medium for multimedia information such as images. However, DNA storage has a low coding density and weak error correction ability.

Results To achieve more efficient DNA storage image reconstruction, we propose DNA-QLC (QRes-VAE and Levenshtein code (LC)), which uses the quantized ResNet VAE (QRes-VAE) model and LC for image compression and DNA sequence error correction, thus improving both the coding density and error correction ability. Experimental results show that the DNA-QLC encoding method can not only obtain DNA sequences that meet the combinatorial constraints, but also have a net information density that is 2.4 times higher than DNA Fountain. Furthermore, at a higher error rate (2%), DNA-QLC achieved image reconstruction with an SSIM value of 0.917.

Conclusions The results indicate that the DNA-QLC encoding scheme guarantees the efficiency and reliability of the DNA storage system and improves the application potential of DNA storage for multimedia information such as images.

Keywords Levenshtein code, Net information density, Combinatorial constraint, Image reconstruction

Introduction

With the rapid development of emerging technologies such as artificial intelligence, big data and blockchain, massive image data continues to emerge, and traditional storage media can no longer meet such huge storage needs. With the advantages of high storage density, long storage time, and low energy consumption, DNA storage has become one of the potential media to solve the future data storage crisis. The quality of DNA encoding and error correction methods will directly affect the cost of synthesis, sequencing and the integrity of data reading and writing, so it has attracted widespread attention from many researchers.

Early encoding methods [1–4] primarily relied on specific mapping rules to convert data into DNA sequences. To minimize the risk of errors in the DNA sequences during storage, it was essential to design DNA sequences that adhered to constraints like GC content and homopolymers. However, this approach often resulted in reduced information density. Subsequent research introduced alternative encoding methods, such as DNA Fountain [5] and the Yin-Yang codec system [6], which explored new solutions without compromising information density. However, the fountain code encoding method requires sufficient redundancy to ensure successful decoding. To address this issue, the Yin-Yang codec system has been proposed, which not only reduces decoding redundancy but also yields highly robust DNA encoding sequences. In addition, in view of the high correlation of image data,

*Correspondence:
Qiang Zhang
zhangq@dlut.edu.cn
[1] School of Computer Science and Technology, Dalian University of Technology, Lingshui Street, DalianLiaoning 116024, China
[2] The Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, School of Software Engineering, Dalian University, Xuefu Street, DalianLiaoning 116622, China

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

a variety of coding methods have emerged. These include coding solutions based on biological constraints [7], lossy/lossless hybrid coding schemes [8], and BO-DNA medical image coding models [9].

Although the aforementioned encoding methods have been designed to maximize information density while meeting GC content and homopolymer constraints in DNA sequences, thereby enhancing the stability and robustness of DNA storage systems, they do not entirely ensure data recovery. This is due to the high susceptibility of DNA storage systems to errors, primarily intra-sequence and inter-sequence errors. Intra-sequence errors typically occur during data writing (synthesis) and reading (sequencing) phases, leading to potential issues such as substitution, deletion, and insertion [10–12]. On the other hand, inter-sequence errors refer to the loss of sequences [13, 14]. In order to deal with this challenge, various error correction methods are designed. Currently, to address sequence loss issues, the primary approach involves adding redundant sequences [5, 15]. In the early stage, error correction codes were mainly used to correct base errors in sequences, such as RS (Reed-Solomon) codes [3, 16–18], BCH (Bose-Chaudhuri-Hocquenghem) codes [19], LDPC (low-density parity-check) codes [20, 21], HEDGES (Hash Encoded, Decoded by Greedy Exhaustive Search) error-correcting code [22], and DNA-Aeon [23]. Besides, there are also methods available that achieve error correction based on specific constraints [24] and particular rules [25]. Given the extensive exploration and research conducted by researchers in the field of encoding and error correction within DNA storage, the current systems still confront issues concerning low encoding density and relatively weak error correction capabilities.

In order to address these challenges, we propose a DNA-QLC encoding scheme. First, to improve the coding density for DNA storage systems, the scheme applies QRes-VAE (for quantized ResNet VAE) to compress images into several bitstreams. After that, the LC (Levenshtein code) is used to add check bits to the bitstreams to realize the correction of substitution, deletion, and insertion errors during DNA storage, which can solve the problem of weak error correction ability. Finally, mapping rules are used to encode bitstreams into the DNA sequence that meet combinational constraints (local GC content of 50% and homopolymers with a size of less than 2) and to avoid the occurrence of undesired motifs (GAATTC and GGC) in DNA sequences, thus improving the robustness of DNA sequences.

Results

This study proposes DNA-QLC to enhance the performance of DNA storage systems by increasing the coding density and error correction capability. DNA-QLC was compared with previous representative works [1–6] in terms of coding results, error correction performance, and synthesis cost. The experimental results show that the DNA sequences encoded by DNA-QLC meet constraints such as local GC content and homopolymers and also avoid the occurrence of two undesired motifs, improving the robustness of DNA sequences. Moreover, given the compression ability achieved by DNA-QLC when storing images, the net information density reached by the scheme is 2.90 bits/nt, reducing the synthesis cost. In particular, at a high error rate, the SSIM value of the image before and after encoding by DNA-QLC is close to 1, indicating that the scheme has excellent error correction performance.

Encoding result

To prove the advantages of the DNA-QLC coding scheme, we compared it with representative coding schemes [1–6] for the same image encoding results. As shown in Table 1, in terms of net information density, DNA-QLC breaks through the limit of 2 bits per base and reaches 2.90 bits/nt, which makes it the maximum among these encoding schemes. In terms of biological constraints, the previous encoding scheme can only

Table 1 Comparison of encoding schemes

Method/Reference	Error correction strategy	Number of oligos	Net information density (bits/nt)	GC content (%)	Maximum homopolymer length (nt)	Avoidance of undesired motifs
Church/ [1]	No	4064	0.94	39–61	3	No
Goldman/ [2]	Repetition	3251	1.48	39–60	1	Yes
Grass/ [3]	RS	2787	1.56	36–62	3	No
Blawat/ [4]	RS	2787	1.40	24–60	3	No
Erlich/ [5]	Fountain	2927	1.23	39–62	4	No
Yin-Yang/ [6]	RS	3125	1.36	40–60	4	No
DNA-QLC	LC	**1293**	**2.90**	**50**	2	**Yes**

maintain the GC content between 40 and 60%, while DNA-QLC can control the local GC content at 50%. Similar to encoding schemes, considering the limitations of biotechnology, DNA-QLC controls the length of the homopolymer within 2 bases, which can significantly decrease the probability of errors in the process of reading and writing. In addition, the DNA-QLC encoding scheme can avoid the occurrence of two undesired motifs (GAATTC and GGC) and reduce the probability of sequence loss and the reading error rate. We used a histogram (Fig. 1) to show the situation of undesired motifs in different encoding schemes. It can be seen more intuitively that only Goldman and DNA-QLC are free from undesired motifs. In sum, besides maintaining a high information density, DNA-QLC makes the DNA sequence more highly adaptable to the process of "writing" and "reading" in the DNA storage system, improving the stability and reliability of DNA storage.

To assess the reconstructed image's quality, we measured the degree of distortion of the image and the similarity of the two images before and after encoding by SSIM [26]. SSIM is a perceptual model that correlates well with the visual experience of human eyes and considers three crucial aspects of an image: luminance, contrast, and structure. The maximum SSIM value is 1, and the minimum SSIM value is −1. A higher SSIM value indicates a higher similarity between the original and the reconstructed image, while a lower value suggests a greater difference between the two images. The data source for SSIM value calculation was the ILSVRC2012 dataset [27], and Fig. 2 displays the results of the image comparison. The figure shows that the image obtained using DNA-QLC is visually similar to the original image and that the main objects in these images are accurately captured. For different images, DNA-QLC has a different SSIM (S) and net information density (N), but both the SSIM and the net information density are competitive. This is because the QRes-VAE model has a certain compression function that enables DNA-QLC to significantly increase the net information density while guaranteeing image quality.

Error correction performance evaluation

To assess how effectively DNA-QLC corrects errors, we compared SSIM values with those of previous representative encoding schemes (Grass [3] and Yin-Yang [6]) under different error rates, as shown in Fig. 3. Three kinds of encoding schemes were used to encode the Mona Lisa image, and then we randomly added three types of errors: substitution, deletion, and insertion. Since substitution errors are more likely to occur than the other two types of errors, definition substitution errors account for half of the total errors when simulating errors. Here, the three encoding schemes were run 10 times under the same error rate (the error rate was maintained at 0.1–2%) to calculate the mean value and standard error. In Fig. 3,

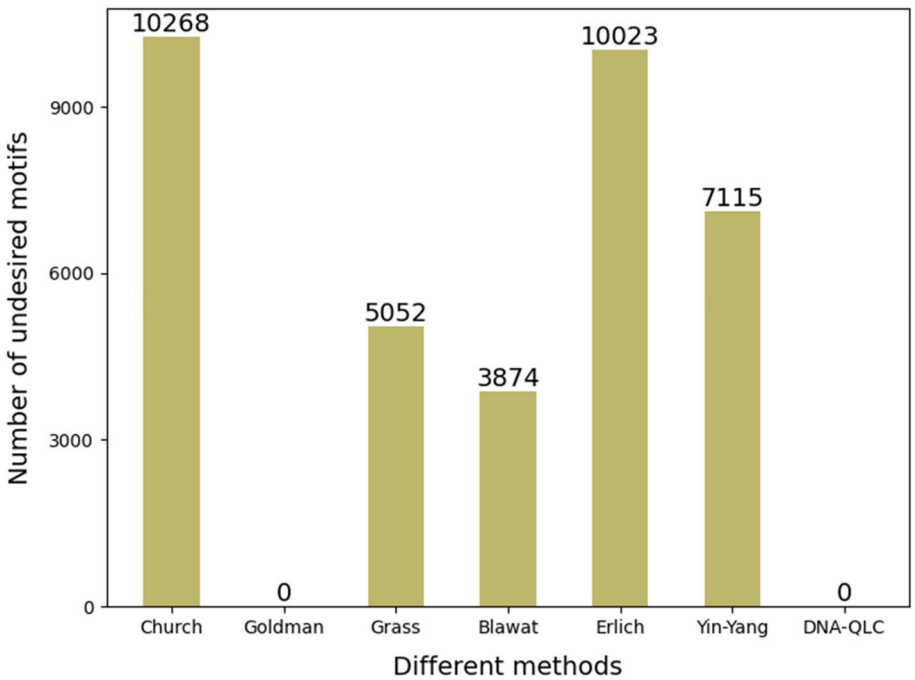

Fig. 1 The case of undesired motifs in each encoding scheme

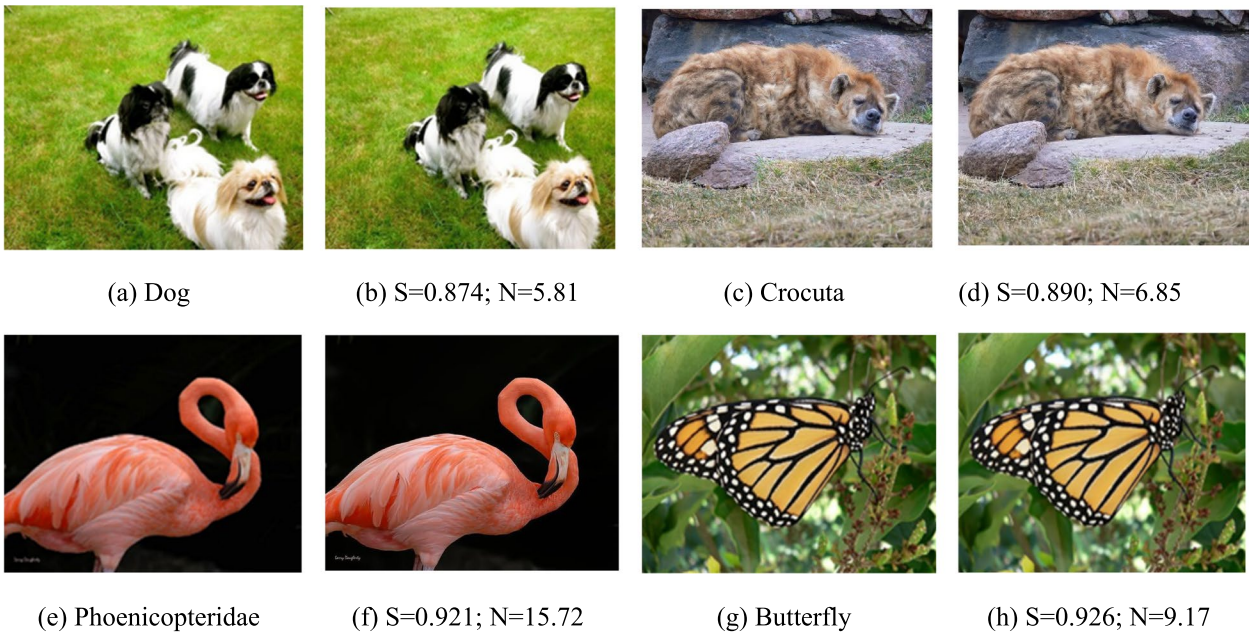

Fig. 2 Graphic view before and after different image encoding. **a**, **c**, **e** and **g** are original image examples. **b**, **d**, **f** and **h** are based on the DNA-QLC with the results of SSIM (S) and Net information density (N, bits/nt)

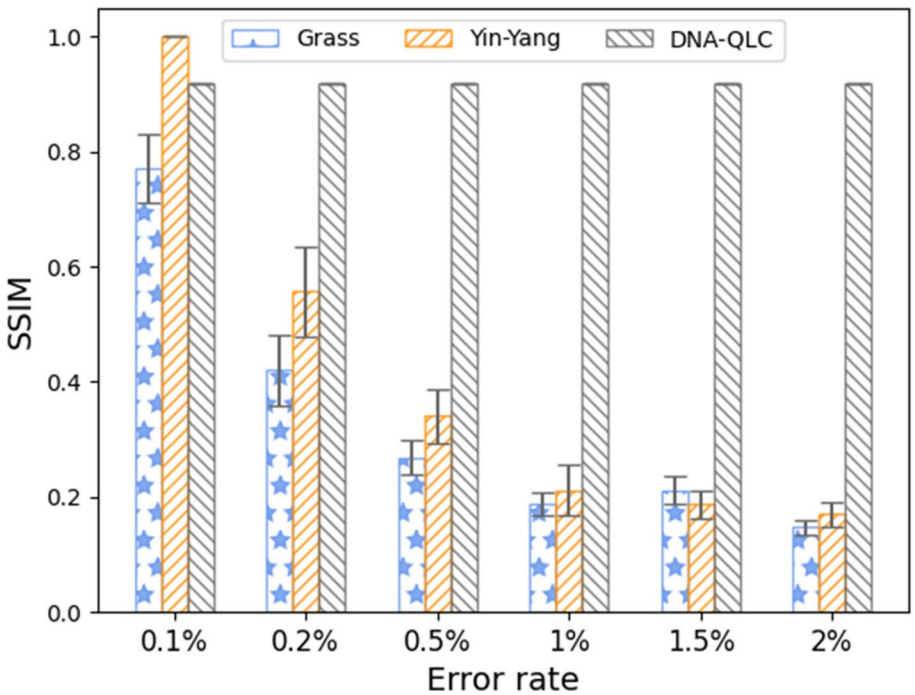

Fig. 3 Comparison of the error correction performance of different encoding schemes under different error rates

when the error rate is 0.1%, the Yin-Yang code has a higher SSIM value, which is equal to the optimal value 1, indicating that the images before and after encoding are completely consistent. However, with the increase in the error rate, the SSIM value of the encoding scheme drops sharply. As the error rate increases, the SSIM value of the Grass encoding scheme decreases clearly. However, the SSIM value of the DNA-QLC encoding scheme remains

stable at 0.917 with the increase in the error rate, indicating that the scheme has a stronger error correction ability in the case of a high error rate. The shorter error bars also exclude sampling errors. Compared with the RS error correction code (Yin-Yang and Grass), DNA-QLC corrects all errors by inserting check bits into the binary stream, and it overcomes the defect that the RS error correction code has a positive correlation between the error correction performance and redundancy. Moreover, the error correction result of DNA-QLC does not fluctuate, because this scheme generates multiple DNA sequence files during the error correction process and then selects a correct sequence file from them to achieve error-free image reconstruction. Figure 4 shows the image reconstructions with error rates of 0.5% and 1.5%. DNA-QLC achieves the best visual results, further illustrating the significant advantages of the encoding scheme's error correction capability.

Costs of synthesizing analysis

Owing to the limitations of current biotechnology, it is expensive to synthesize DNA sequences. Therefore, the encoding scheme should achieve a good error correction performance and also improve the utilization rate of the base to reduce the synthesis cost. To assess the cost, we compared DNA-QLC with four published open-source encoding schemes, including DNA Fountain [5] and Yin-Yang [6]. The Mona Lisa image data were encoded using five encoding schemes, and the oligonucleotide pool pricing table from Twist Bioscience [28] was used to approximate the synthesis cost of the encoding sequence, as shown in Fig. 5. Church's encoding scheme requires 4064 sequences of 204 bases each, for a total of 829,056 bases to decode to obtain the Mona Lisa image. Grass' encoding scheme requires 2787 sequences of 180 bases each. DNA Fountain requires 2927 sequences, each with 216 bases. Yin-Yang requires 3125 sequences of 184 bases each. DNA-QLC requires 1293 sequences of 208 bases each. As can be seen from Fig. 5, when a 95.2-KB image (Mona Lisa.jpg) is stored, the synthesis cost required by the DNA-QLC encoding scheme is the lowest. The DNA-QLC is 33.3% cheaper than the most widely used DNA Fountain encoding scheme and 11.6% cheaper than the latest Yin-Yang encoding scheme.

Conclusion

In this study, aiming at the problems of low coding density and weak error-correcting ability in DNA storage, we proposed a DNA-QLC encoding scheme that uses the QRes-VAE model and the LC algorithm

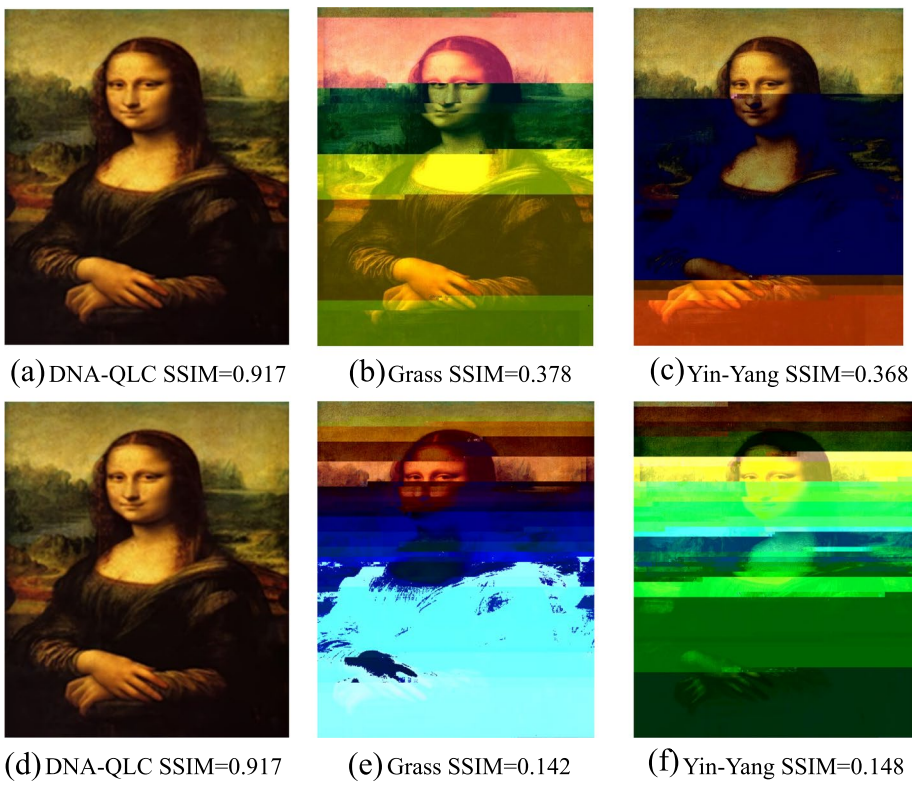

Fig. 4 Graphic view of one example. **a**, **b** and **c** are the reconstruction of the image when the error rate of the three coding schemes is 0.5%. **e**, **f** and **g** are the reconstruction of the image when the error rate of the three coding schemes is 1.5%

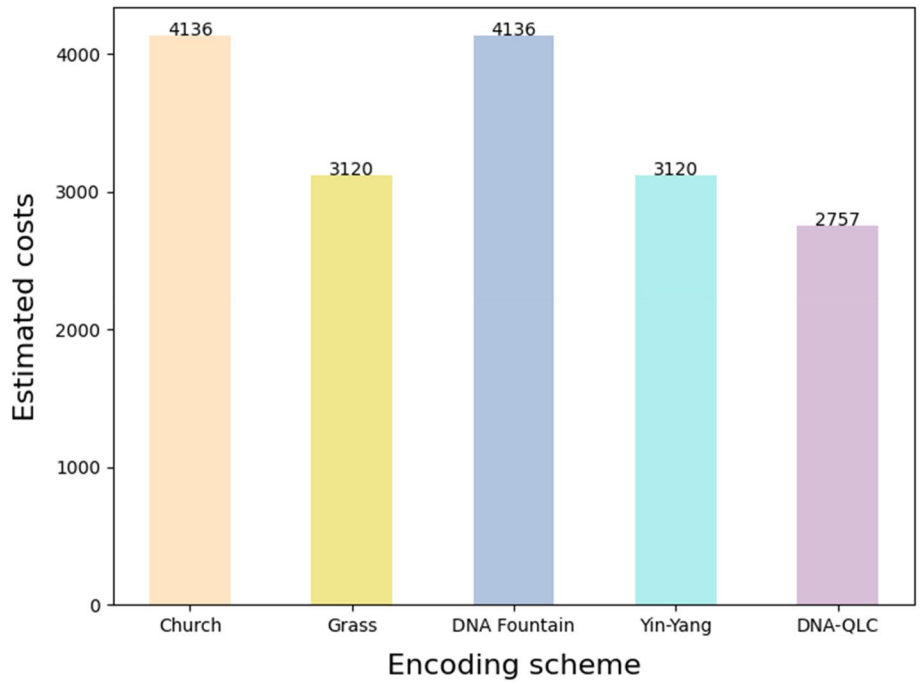

Fig. 5 Cost evaluation of different encoding schemes

to compress images and correct mistakes. Comparing DNA-QLC with representative encoding schemes encoding the same image, the net information density reached by DNA-QLC is 2.90 bits/nt, 2.4 times that of DNA fountain codes (Table 1). The results in Fig. 2 show that when the input image contains significant amounts of redundant information, the net information density of DNA-QLC is 15.72 bits/nt. Clearly, the introduction of the QRes-VAE model can significantly improve the encoding density, greatly reducing the cost of DNA storage. The DNA-QLC only uses simple mapping rules to encode bitstreams into DNA sequences that meet the local GC content level of 50%, the homopolymer length of no more than 2, and no undesired motifs, effectively reducing the DNA probability of errors during DNA storage. In addition, DNA-QLC can also detect and correct multiple errors of the same type through the LC. Based on the experimental findings, we can conclude that the DNA-QLC encoding scheme can overcome the problem of the positive correlation between the error correction performance and the redundancy of other encoding schemes. And with the increase of the error rate, the image SSIM value will not decrease. In addition, when the error rate is high, the DNA-QLC encoding scheme can still maintain the integrity and clarity of the image, and they do not cause serious distortion or the failure to recognize the main object (Fig. 4).

Although DNA-QLC has high encoding and error correction performance, DNA-QLC has a defect in correcting the substitution error, that is, it can only correct the errors of purine mutation to pyrimidine or pyrimidine mutation to purine. We will attempt to resolve this issue in future research work as well as study the molecular characteristics of DNA, construct a DNA storage self-error correction model based on deep learning, reduce the overhead of unnecessary error correction redundancy, further improve the net information density and capacity of DNA storage, reduce its cost, and promote DNA storage practical applications in storing cold data.

Methods

To improve the coding density and error correction performance in the DNA storage system, we propose a DNA-QLC encoding scheme, which is primarily divided into QRes-VAE compression model and LC encoding algorithm. First, the input image is compressed using the QRes-VAE model to obtain compressed binary data, which are segmented and indexed. Then, the LC is used to add check bits to the binary data. Finally, the bitstreams are mapped to DNA sequences that comply with the combinatorial constraints through mapping rules. The flowchart and pseudocode of DNA-QLC are shown in Fig. 6 and Algorithm.

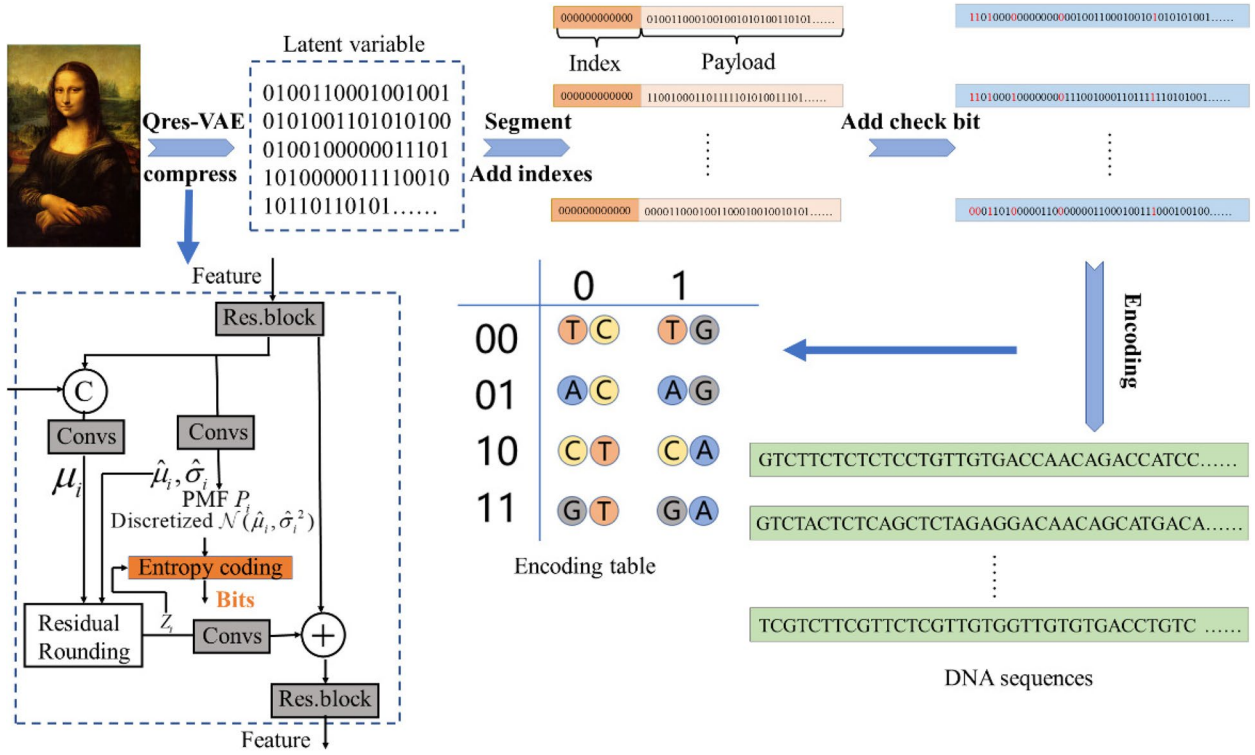

Fig. 6 Flowchart of the DNA-QLC encoding scheme

Algorithm 1. The pseudocode of DNA-QLC

```
Input: The images
Output: The reconstructed image
1 Use the QRes-VAE model to extract the potential variable.
2 Split binary data and add indexes.
3 Add check bits to the binary fragment.
4 Convert binary data into DNA sequences.
5 Correct the error:
    for i in range (len(sequences)):
            if l==n
               if base is substituted
             sequence = subcreat(sequences[i], bases, index)
          else
            break
        elif l<n
          sequence =delcreat(sequences[i], sequences, n)
        else:
          sequence = inscreat (sequences[i], sequences, n)
6 Convert the corrected DNA sequence into binary data.
7 Remove check bits.
8 Sort the binary data by index and then merge the data.
9 Obtain the reconstructed image through the QRes-VAE model.
```

Image compression with QRes-VAE

Compression technology is critical for DNA storage. An efficient compression method can improve the utilization rate of the base while ensuring the integrity of data, increasing the coding density, and reducing the synthesis cost. However, the performance of current compression techniques in DNA storage is lacking, such as Huffman coding [2] and discrete wavelet decomposition [7]. In this paper, images are compressed by the QRes-VAE model, and the compression process of this model [29] is described in detail in this subsection.

Network architecture

The QRes-VAE model is based on ResNet (residual network) VAE [30], using quantization-aware posteriors and priors to redesign latent variables. It consists of a bottom-up path and a top-down path. When an image is inputted, the bottom-up path produces some deterministic features, which are then transmitted to the top-down path for inference, after which the image is reconstructed by up-sampling.

Loss function

The loss function of the QRes-VAE model [29] is shown as follows:

$$\begin{aligned}\mathcal{L} &= D_{\mathrm{KL}}\left(q_{Z|x}\,\|p_z\right) + \mathbb{E}_{q_{Z|x}}\left[\log\tfrac{1}{P_{X|Z}(x|Z)}\right],\\ &= \mathbb{E}_{q_{Z|x}}\left[\sum_{i=l}^{N}\log\tfrac{1}{P_i(Z_i|Z_{<i})} + \lambda\cdot d\left(x,\widehat{x}\right)\right] + \text{constant},\end{aligned} \tag{1}$$

where $Z \sim q_{Z|x}$ represents the estimation of samples extracted in each training step, λ is the hyperparameter that can be set manually, x is the input image, $\overline{x}$ is the reconstruction image, and $d(x,\overline{x})$ is the mean square error of the input image and reconstruction image.

Compression

First, features are extracted from the input image and quantized into N potential variables $(Z_1, Z_2, \ldots, Z_N)$. Then, potential variables are encoded into bits by using the probability quality function (PMF) [31]. Finally, entropy encoding is performed by the range-based asymmetric numeral systems (rANS) [32]. The quantization formula [29] during compression is as follows:

$$z \leftarrow \widehat{v}_i + \lfloor v_i - \widehat{v}_i \rceil, \tag{2}$$

where $\lfloor\cdot\rceil$ is the nearest integer function, and v_i is quantified as its nearest neighbor in the set $\left\{\widehat{v}_i + n | n \in \mathbb{Z}\right\}$, denoted by Z_i.

The formula for the PMF $P_i(\cdot)$ [29] is given below.

$$P_i(n) \triangleq p_i\left(\widehat{v}_i + n | Z_{<i}\right), n \in \mathbb{Z} \tag{3}$$

Decompression is the inverse process of compression. It first uses rANS to decode each bitstream and transform Z_i using convolution layers before adding it to the feature. Finally, it obtains the reconstructed image through the final up-sampling layer in the top-down decoder. The QRes-VAE model is used to compress images so that a small number of DNA sequences can be used to store image information. This approach improves the coding density of the DNA storage system and reduces the cost.

Levenshtein code algorithm

In the process of DNA synthesis, PCR amplification and DNA sequencing, substitution, deletion and insertion errors are easy to occur. Previous studies have reported that chemistry synthesis and second-generation sequencing result in an error rate of about 1% per base [15] and that third-generation sequencing has an error rate of up to 10% [33]. Moreover, the error rate of DNA sequences varies among different motifs. DNA sequences with homopolymers and abnormal GC content are difficult to synthesize [6], thus generally having a higher error rate during the synthesize process. Therefore, constraints are crucial to avoiding errors in DNA storage, and more new constraints are being widely explored and studied. For example, DNA sequences with restriction sites were easily cleaved by restriction enzymes ("GAATTC" for "EcoRI") during in vivo storage, leading to information loss [34]. For the replication process, local GC content balance was explored to improve the success rate of PCR amplification technology [35]. During the sequencing process, the Illumina sequencing platform had a higher error probability for DNA sequences containing "GGC" fragments [36].

Most current encoding schemes can only meet the two basic constraints of a global GC content and homopolymer and cannot satisfy the abovementioned new constraints. Therefore, a novel mapping rules was designed, whose central idea is to map three binary numbers to two bases (e.g., 000 → TC, 001 → TG, 010 → AC, 011 → AG, 100 → CT, 101 → CA, 110 → GT, 111 → GA). The mapping rules, which sets purine and pyrimidine in series, can well control the local GC content to 50% and simultaneously set keep the maximum limit of homopolymers 2 and exclude the occurrence of two undesired motifs (GAATTC and GGC). However, constraint encoding can only reduce the error probability, but can not completely avoid it. To further ensure the read–write integrity of data, LC [37] is used to correct substitution, deletion, and insertion errors that occur within the sequence.

LC

This is a binary algebraic code whose binary codeword of length n satisfies Eq. (4) [37].

$$L(m,r,U)=\left\{x\in\{0,1\}^m:\sum_{k=1}^{m}x_k*k\equiv r \bmod U\right\} \tag{4}$$

For any integer $U\geq 2m$, $0\leq r\leq U-1$, in this study, let $r=0$ and $U=2m$. In sum, an l bits binary sequence is processed into a codeword of length m, and the conditions below should be satisfied.

$$\sum_{k=1}^{m}x_k*k\equiv 0 \bmod 2m \tag{5}$$

The central idea of the LC is actually to insert parity bits at 2^i-th positions to ensure that the codeword has a desired syndrome. Note that the last position is always a parity bit and that the second-to-last position is the message bit. When the length of the binary data is l, the length of the codeword processed by the is m, and the calculation equation is as below.

$$l=m'-\lceil\log_2 m'\rceil-1 \tag{6}$$

$$m=\min m' \tag{7}$$

Example

The binary data are 101,000,011,110, and the length $l=12$, which is calculated by Eqs. (6) and (7) to obtain $U=36$. To meet the conditions of Eq. (5), we must use an additional 6 check bits.

Calculate the syndrome of the bits using $\sum_{k=1}^{m}x_k*k=3+6+12+13+14+15=63$. Through $2U-63$, the syndrome can be calculated to be equal to 9, we can convert it into binary (001001) and obtain 6 check bits, then we insert that value into the binary data stream and obtain the code word 101,001,010,001,111,000 processed by LC. Then, according to the mapping rules, the code word 101,001,010,001,111,000 can be encoded into the DNA sequence CATGACTGGATC, which satisfies the local GC content of 50%, the homopolymer length not exceeding 2, and no GAATTC and GGC two undesired motifs. In addition, if the sequence has substitution, deletion, and insertion errors during the DNA storage process, the added syndrome can be used to correct the sequence to obtain the correct original information.

Acknowledgements
We thank our partners who provided all the help during the research process and the team for their great support.

Authors' contributions
Y.Z. designed the study, drafted the manuscript. X.Z. and S.C. provided knowledge guidance. B.C., B.W. and Q.Z. supervised the experiment and revised the manuscript. All authors read and approved the final manuscript.

Funding
This work is supported by 111 Project (No. D23006), the National Natural Science Foundation of China (Nos. 62272079, 61972266), Natural Science Foundation of Liaoning Province (No. 2022-KF-12-14), the Postgraduate Education Reform Project of Liaoning Province (No. LNYJG2022493), the Artificial Intelligence Innovation Development Plan Project of Liaoning Province (No. 2023JH26/10300025), the Dalian Outstanding Young Science and Technology Talent Support Program (No. 2022RJ08), Dalian Major Projects of Basic Research (No. 2023JJ11CG002).

Availability of data and materials
The data and code underlying this article are available in https://github.com/Larissa-11/DNA-QLC.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 June 2023 Accepted: 1 March 2024
Published online: 09 March 2024

References

1. Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science. 2012;337(6102):1628–1628.
2. Goldman N, Bertone P, Chen SY, Dessimoz C, LeProust EM, Sipos B, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature. 2013;494(7435):77–80.
3. Grass RN, Heckel R, Puddu M, Paunescu D, Stark WJ. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew Chem Int Ed Engl. 2015;54(8):2552–5.
4. Blawat M, Gaedke K, Huetter I, Chen X-M, Turczyk B, Inverso S, et al. Forward error correction for DNA data storage. Procedia Comput Sci. 2016;80:1011–22.
5. Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture. Science. 2017;355(6328):950–3.
6. Ping Z, Chen S, Zhou G, Huang X, Zhu SJ, Zhang H, et al. Towards practical and robust DNA-based data archiving using the yin–yang codec system. Nat Comput Sci. 2022;2(4):234–42.
7. Dimopoulou M, Antonini M, Barbry P, Appuswamy R. A biologically constrained encoding solution for long-term storage of images onto synthetic DNA. 2019 27th European Signal Processing Conference (EUSIPCO). 2019. p. 1–5.
8. Li Y, Du DH, Ou L, Li B. HL-DNA: A hybrid lossy/lossless encoding scheme to enhance DNA storage density and robustness for images. 2022 IEEE 40th International Conference on Computer Design (ICCD). 2022. p. 434–42.
9. Rasool A, Hong J, Jiang Q, Chen H, Qu Q. BO-DNA: Biologically optimized encoding model for a highly-reliable DNA data storage. Comput Biol Med. 2023;165:107404.
10. Dong Y, Sun F, Ping Z, Ouyang Q, Qian L. DNA storage: research landscape and future prospects. Natl Sci Rev. 2020;7(6):1092–107.
11. Zheng Y, Cao B, Wu J, Wang B, Zhang Q. High net information density DNA data storage by the MOPE encoding algorithm. IEEE/ACM Trans Comput Biol Bioinform. 2023;20(5):2992–3000.
12. Heckel R, Mikutis G, Grass RN. A characterization of the DNA data storage channel. Sci Rep. 2019;9(1):1–12.
13. Chen KK, Zhu JB, Boskovic F, Keyser UF. Nanopore-Based DNA Hard Drives for Rewritable and Secure Data Storage. Nano Lett. 2020;20(5):3754–60.

14. Gao YM, Chen X, Qiao HY, Ke YG, Qi H. Low-bias manipulation of DNA oligo pool for robust data storage. ACS Synth Biol. 2020;9(12):3344–52.
15. Bornhol J, Lopez R, Carmean DM, Ceze L, Seelig G, Strauss K. A DNA-based archival storage system. ACM Sigplan Not. 2016;51(4):637–49.
16. Meiser LC, Antkowiak PL, Koch J, Chen WD, Kohll AX, Stark WJ, et al. Reading and writing digital data in DNA. Nat Protoc. 2020;15(1):86–101.
17. Cao B, Zhang X, Cui S, Zhang Q. Adaptive coding for DNA storage with high storage density and low coverage. NPJ Syst Biol Appl. 2022;8(1):23.
18. Yan Z, Liang C, Wu H. A segmented-edit error-correcting code with re-synchronization function for DNA-based storage systems. IEEE Trans Emerg Top. 2022;11(3):605–18.
19. Chen WG, Wang LX, Han MZ, Han CC, Li BZ. Sequencing barcode construction and identification methods based on block error-correction codes. Sci China Life Sci. 2020;63(10):1580–92.
20. Chen WG, Han MZ, Zhou JT, Ge Q, Wang PP, Zhang XC, et al. An artificial chromosome for data storage. Natl Sci Rev. 2021;8(5):nwab028.
21. Lenz A, Maarouf I, Welter L, Wachter-Zeh A, Rosnes E, i Amat AG. Concatenated codes for recovery from multiple reads of DNA sequences, 2020 IEEE Information Theory Workshop (ITW). 2021. p. 1–5.
22. Press WH, Hawkins JA, Jones SK Jr, Schaub JM, Finkelstein IJ. HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints. Proc Natl Acad Sci. 2020;117(31):18489–96.
23. Welzel M, Schwarz PM, Löchel HF, Kabdullayeva T, Clemens S, Becker A, et al. DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in DNA storage. Nat Commun. 2023;14(1):628.
24. Cai K, Chee YM, Gabrys R, Kiah HM, Nguyen TT. Correcting a single indel/edit for DNA-based data storage: linear-time encoders and order-optimality. IEEE Trans Inf Theory. 2021;67(6):3438–51.
25. Li XY, Chen MX, Wu HM. Multiple errors correction for position-limited DNA sequences with GC balance and no homopolymer for DNA-based data storage. Brief Bioinform. 2023;24(1):bbac484.
26. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12.
27. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.
28. What is the typical pricing for oligo pools? https://www.twistbioscience.com/faq/oligo-pools/what-typical-pricing-oligo-pools. Accessed 31 May 2023.
29. Duan Z, Lu M, Ma Z, Zhu F. Lossy image compression with quantized hierarchical VAEs. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023. p. 198–207.
30. Vahdat A, Kautz J. NVAE: A deep hierarchical variational autoencoder. Adv Neural Inf Process. 2020;33:19667–79.
31. Ballé J, Minnen D, Singh S, Hwang SJ, Johnston N. Variational image compression with a scale hyperprior. arXiv. 2018.
32. Duda J. Asymmetric numeral systems: entropy coding combining speed of Huffman coding with compression rate of arithmetic coding. arXiv. 2013;1311:2540.
33. Ping Z, Ma DZ, Huan XL, Chen SH, Liu LY, Guo F, et al. Carbon-based archiving: current progress and future prospects of DNA-based data storage. Gigascience. 2019;8(6):giz075.
34. Polisky B, Greene P, Garfin DE, McCarthy BJ, Goodman HM, Boyer HW. Specificity of substrate recognition by the EcoRI restriction endonuclease. Proc Natl Acad Sci. 1975;72(9):3310–4.
35. Gabrys R, Kiah HM, Vardy A, Yaakobi E, Zhang Y. Locally balanced constraints. 2020 IEEE International Symposium on Information Theory (ISIT). 2020. p. 664–9.
36. Bornholt J, Lopez R, Carmean DM, Ceze L, Seelig G, Strauss K. Toward a DNA-based archival storage system. IEEE Micro. 2017;37(3):98–104.
37. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady. 1966;10(8):707–10.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.