small methods
www.small-methods.com

DNA-Based Concatenated Encoding System for High-Reliability and High-Density Data Storage

用于高可靠性和高密度数据存储的基于 DNA 的级联编码系统

Yubin Ren, Yi Zhang, Yawei Liu, Qinglin Wu, Juanjuan Su, Fan Wang,* Dong Chen,* Chunhai Fan, Kai Liu,* and Hongjie Zhang*

Information storage based on DNA molecules provides a promising solution with advantages of low-energy consumption, high storage efficiency, and long lifespan. However, there are only four natural nucleotides and DNA storage is thus limited by 2 bits per nucleotide. Here, artificial nucleotides into DNA data storage to achieve higher coding efficiency than 2 bits per nucleotide is introduced. To accommodate the characteristics of DNA synthesis and sequencing, two high-reliability encoding systems suitable for four, six, and eight nucleotides, i.e., the RaptorQ-Arithmetic-LZW-RS (RALR) and RaptorQ-Arithmetic-Base64-RS (RABR) systems, are developed. The two concatenated encoding systems realize the advantages of correcting DNA sequence losses, correcting errors within DNA sequences, reducing homopolymers, and controlling specific nucleotide contents. The average coding efficiencies with error correction and without arithmetic compression by the RALR system using four, six, and eight nucleotides reach 1.27, 1.61, and 1.85 bits per nucleotide, respectively. While the average coding efficiencies by the RABR system are up to 1.50, 2.00, and 2.35 bits per nucleotide, respectively. The coding efficiency, versatility, and tunability of the developed artificial DNA systems might provide significant guidance for high-reliability and high-density data storage.

1. Introduction

The unmet need between exponentially increasing data and limited capacity of current mainstream storage media is becoming increasingly prominent.[1–3] Traditional storage media, such as magnetic, optical, and solid-state media, only possess binary coding of "1" or "0" on each bit, and their storage capacity is about to reach its maximum. Thus, it is urgent to develop novel storage media with low-power consumption, high storage density, and long lifespan. Deoxyribonucleic acid (DNA), known as the genetic information carrier,[4,5] has proved to be a promising data storage medium due to its long lifespan,[6–10] sky-high storage density, low energy consumption, and low maintenance cost.[11–13] Natural DNA has four nucleotides (A, T, C, and G) and has a storage density of ≈460 EB g^{-1},[14] which is much higher than that of traditional storage media. Therefore, the development of new types of DNA molecules to improve information storage is an attractive goal.

DNA data storage is mainly affected by the number of nucleotides, the biochemical properties of DNA, and the technical constraints of DNA synthesis and sequencing.[15] Therefore, encoding systems including different functional modules, such as index design and error correction, are required to accommodate the characteristics of the DNA storage channel: i) DNA sequences with acceptable synthetic errors are generally limited to about 250 nucleotides and large data is thus divided into short DNA sequences. Therefore, redundant nucleotides are required to

Y. Ren, K. Liu, H. Zhang
Department of Chemistry
Tsinghua University
Beijing 100084, China
E-mail: kailiu@tsinghua.edu.cn

Y. Zhang, Y. Liu, F. Wang
State Key Laboratory of Rare Earth Resource Utilization
Changchun Institute of Applied Chemistry
Chinese Academy of Sciences
Changchun, Jilin 130022, China
E-mail: wangfan@ciac.ac.cn

Q. Wu, D. Chen
Institute of Process Equipment
College of Energy Engineering and State Key Laboratory of Fluid Power and Mechatronic Systems
Zhejiang University
Hangzhou, Zhejiang 310027, China
E-mail: chen_dong@zju.edu.cn

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/smtd.202101335.

DOI: 10.1002/smtd.202101335

J. Su
College of Materials Science and Opto-Electronic Technology
University of Chinese Academy of Sciences
Beijing 100049, China
E-mail: sujuanjuan@ucas.ac.cn

C. Fan
Frontiers Science Center for Transformative Molecules
School of Chemistry and Chemical Engineering
and Institute of Molecular Medicine
Renji Hospital
School of Medicine
Shanghai Jiao Tong University
Shanghai 200240, China

index the DNA sequences for data reconstruction[13,16] and redundant sequences are required for error correction caused by DNA sequence losses. ii) The existence of homopolymers with repeating bases ≥3 and inappropriate GC contents may increase the error rate in DNA sequencing.[17] Therefore, DNA homopolymers should to be avoided and GC content should to be controlled in a low ratio. iii) Limited by DNA techniques, nucleotide insertion, deletion, and substitution within DNA sequences may occur during the process of DNA synthesis, storage, and sequencing.[18,19] Therefore, redundant nucleotides are required to correct errors within DNA sequences. This may also lead to a high capability of correcting more errors in the DNA sequences. However, extra redundant nucleotides will reduce the coding efficiency. Therefore, the development of new encoding systems with appropriate redundant nucleotides is becoming important to ensure both high reliability and high coding efficiency for DNA data storage.

To address the problems, a lot of effort has been dedicated to develop encoding algorithms suitable for DNA data storage. i) Unlike electronic storage media, encoded DNA sequences do not have structured addressing and are stored in a stochastic manner.[20] A fixed number of nucleotides are generally used to index the address of each DNA sequence. ii) To control GC content and homopolymer, different strategies are developed. For example, an encoding algorithm that uses two nucleotides representing a single binary bit, i.e., A or C = 0 and G or T = 1, is developed, in which alternative nucleotides can be used;[21] ternary cyclic encoding with preorder constraint maps the base-3 Huffman code to one of the three nucleotides different from the previous one has been used.[22,23] Each symbol in Galois Field with 48 symbols is mapped to a string of three nucleotides in DNA codon wheel, in which the nucleotides at the second and third positions are different.[7] iii) To ensure the reliability of data storage, the design of DNA data storage algorithms tends to use logical redundancy instead of replica redundancy. Linear Block Codes could correct errors in information transmission by adding check symbols (logical redundancy) and Reed-Solomon (RS) code, a kind of cyclic code in Linear Block Codes, can correct random errors and burst errors in data transmission while keeping a relatively low redundancy.[7,24–26] iv) Unlike linear block code, fountain code could encode information into multiple DNA sequences and could retract the information from a certain number of random DNA sequences.[27] LT digital fountain code is applied to the DNA data storage and achieves a decoding failure rate < 10^{-8} with only <5% redundancy.[28] RaptorQ fountain code, which has higher coding efficiency and reliability than LT fountain code, is also developed.[27] v) Huffman encoding, a lossless compression algorithm, is widely used in DNA data storage to reduce redundancy and improve coding efficiency.[22,29] vi) DNA only has four natural nucleotides, i.e., A, T, C, and G, which limit the maximum coding efficiency to 2 bits per nucleotide. Recently, the number of nucleotides is extended to eight nucleotides, i.e., A, T, C, G, P, Z, S, and B, which form four orthogonal pairs.[30] The introduction of four artificial nucleotides has the potential of realization of coding efficiency to 3 bits per nucleotide. Despite the advances in DNA encoding, there still lacks a systemic encoding, which addresses the shortcomings of DNA storage. Thus, to develop an optimal encoding system, which ensures high coding efficiency and reliability for DNA storage, is urgently needed.

Here, the factors that affect data reliability and coding efficiency are analyzed in detail and two practical high-reliability DNA encoding systems are developed. The two encoding systems include tandem encoders of RaptorQ encoding for correcting DNA sequence loss, arithmetic encoding for reducing redundancy, improved Base64 encoding, or improved Lempel–Ziv–Welch (LZW) encoding for controlling homopolymer and specific nucleotide content, and RS code for correcting errors within DNA sequence. The RaptorQ-Arithmetic-Base64-RS (RABR) and RaptorQ-Arithmetic-LZW-RS (RALR) encoding systems could ensure a high coding efficiency and reliability despite sequence loss and readout error encountered in the DNA data storage channel. The two encoding systems could be extended from four nucleotides to six and eight nucleotides containing both natural and artificial nucleotides and are applicable to text, picture, and video. The absence of arithmetic compression in the coding systems facilitates an objective evaluation of the coding efficiency. The average coding efficiencies with error correction and without arithmetic compression by the RALR system using four, six, and eight nucleotides reach 1.27, 1.61, and 1.85 bits per nucleotide, respectively. While the average coding efficiencies by the RABR system using four, six, and eight nucleotides are up to 1.50, 2.00, and 2.35 bits per nucleotide, respectively. The two encoding systems are versatile and suitable for the DNA data storage channel, providing a systematic scheme for DNA data encoding.

2. Experimental Section

The flow chart of the encoding system, which could be divided into five steps, is shown in Figure 2c. i) A digital file is first divided into binary source blocks, each of which is composed of several source symbols and is independently encoded by the RaptorQ encoder. Intermediate symbols are generated by precoding and repair symbols are generated from intermediate symbols and tuples, which are able to recover lost encoding symbols. ii) After RaptorQ encoding, each block composed of source symbols and repair symbols is independently compressed symbol by symbol using the Arithmetic encoder. iii) Arithmetic symbols are then encoded into DNA symbols by improved Base64 encoder (RABR) or improved LZW encoder (RALR) to obtain DNA symbol sequences with controlled homopolymers and specific nucleotide contents. iv) Index generation algorithm is used to add index symbols at the front of each DNA symbol sequence for later reorganizing the whole source file. v) DNA symbol sequences are converted into binary symbol sequences, each of which is encoded by RS encoder to add controlled redundancy for error correction. After RS encoding, binary symbol sequences are converted back into DNA symbol sequences, which are ready to guide the synthesis of DNA sequences. The original digital file is finally recovered by decoding the readout of DNA sequences. Therefore, the basic encoding process of the RABR system includes systematic RaptorQ encoder, Arithmetic encoder, improved Base64 encoder, and RS encoder, while improved Base64 encoder is replaced by improved LZW encoder in the RALR system.

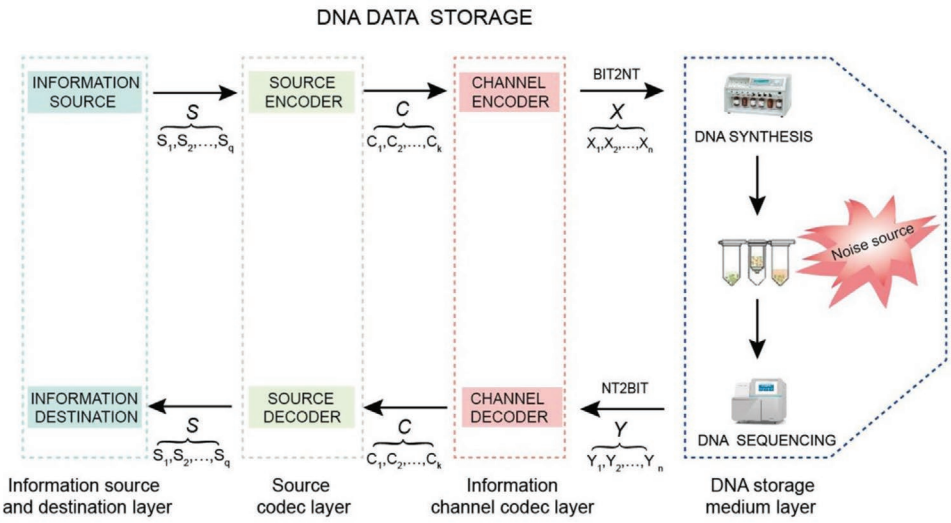

Figure 1. Schematics of the whole process of DNA data storage, including source encoder, channel encoder, DNA synthesis, DNA sequencing, channel decoder and source decoder. These processes could be divided into information source and destination layer, source codec layer, information channel codec layer, and DNA storage medium layer.

3. Results and Discussion

To obtain an in-depth insight into DNA data storage, DNA data storage channel,[18] which includes source encoder, channel encoder, DNA synthesis, DNA sequencing, channel decoder, and source decoder, is demonstrated in **Figure 1**. The whole channel could further be divided into four layers, including information source and destination layer, source codec layer, information channel codec layer and DNA storage medium layer. Because DNA data storage channel is a channel with interference and memory, abstracting strongly correlated symbols into one symbol could simplify it into a discrete memoryless channel. Therefore, an information source is simplified into a discrete memoryless source, $[S, P(s)]$, $s = (s_1, s_2, \ldots, s_q)$, here s_i represents symbols with strong local correlation and $s_i \in \{a_1, a_2, \ldots, a_m\}$, $i = (1, 2, \cdots q)$, while DNA data storage channel is simplified into a discrete memoryless information channel $[X, P(y|x), Y]$: $x = (x_1, x_2, \ldots, x_N), y = (y_1, y_2, \ldots, y_N)$, where x_i and y_i represent symbols with strong local correlation and $x_i \in \{b_1, b_2, \ldots, b_r\}$, $y_i \in \{c_1, c_2 \ldots, c_s\}$, $i = (1, \cdots, N)$, $P(y|x) = P(y_1 y_2 \ldots y_N \mid x_1 x_2 \ldots x_N) = \prod_{i=1}^{N} P(y_i \mid x_i)$ *where* $\sum_y P(y \mid x) = 1$. Generally, the information entropy of a discrete memoryless source is the compressing limit of lossless source coding. Lossless source coding aims to transform source symbols into symbols with equal probability to improve coding efficiency, since the information entropy reaches its maximum when the distribution of source symbols has equal probability. When the information transmission rate $R = I(X; Y)$ is less than the channel capacity $C = \max_{P(x)}\{I(X;Y)\}$, the average error rate of decoding can be considered to be arbitrarily small on the condition that the code length is large enough (Shannon's second theorem). Therefore, a DNA sequence with enough redundancy for error correction is able to ensure the reliability of the DNA data storage channel, while the introduced redundancy will also reduce the coding efficiency at the same time. Thus, it is important to ensure reliability while optimizing the coding efficiency and it is necessary to develop an encoding system, which could obtain the optimal coding efficiency on the premise of ensuring the reliability of DNA data storage. This could only be achieved by analyzing the DNA storage techniques, including synthesis, storage, and sequencing, and developing a suitable encoding system, which could accommodate the characteristics of DNA storage techniques.

The chemical structures of four natural nucleotides, A, T, C, and G, and four artificial nucleotides, P, Z, S, and B, are shown in **Figure 2**a. A–T, C–G, P–Z, and S–B form four orthogonal pairs. The developed encoding system is able to encode text, pictures and videos using the eight nucleotides with a coding efficiency of 2.59, 2.52, and 2.59 bits per nucleotide, respectively, as shown in Figure 2b. The basic encoding process of the RABR and RALR systems is developed based on the characteristics of the DNA data storage channel. Different from conventional storage media, DNA storage writes data by synthesizing DNA sequences, such as solid-phase synthesis, and readouts data by DNA sequencing, such as second-generation sequencing or nanopore sequencing.[31] Therefore, substitution, insertion, and deletion errors in DNA sequences, which may arise during the writing and reading processes, and losses of DNA sequences, which may happen during the storage period, make DNA data storage a great challenge: i) Synthesis of DNA sequences with acceptable synthetic errors is generally limited to about 250 nucleotides. Splitting large data into multiple DNA sequences with controlled length and adding redundant sequences for correcting losses of DNA sequences is achieved by RaptorQ encoder (Supporting Information 1 and Figure S1, Supporting Information). ii) Reducing redundancy is important for increasing the coding efficiency and facilitating the synthesis of DNA sequences. Data compression at symbol level using Arithmetic encoder is performed to reduce the redundancy after RaptorQ encoding (Supporting Information 2 and Figure S2, Supporting Information). iii) The existence of homopolymers and imbalanced contents of specific nucleotides will increase the error rates in DNA sequencing.[32] Homopolymers and specific nucleotide contents are controlled

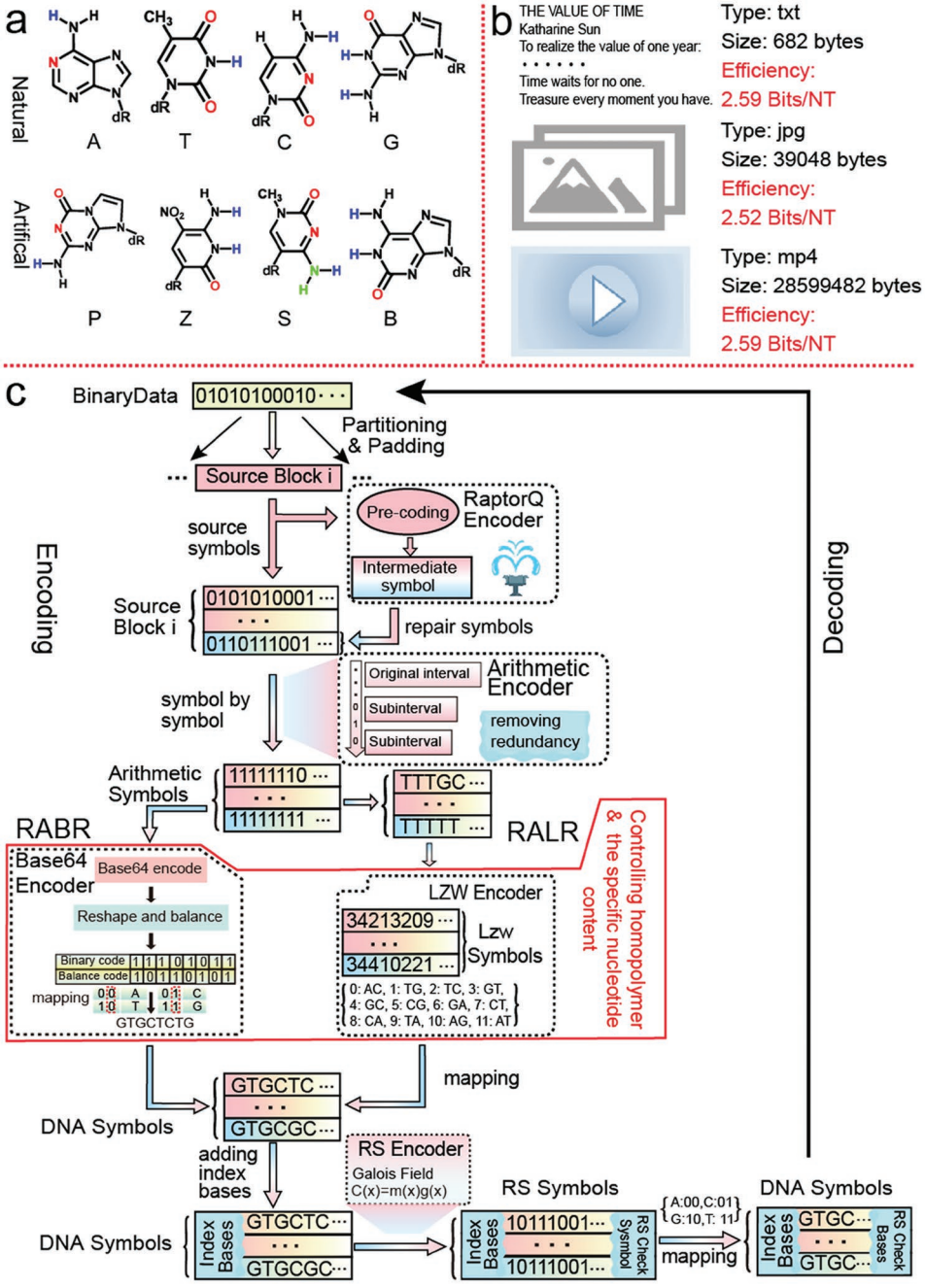

Figure 2. High-density DNA data storage based on both natural and artificial nucleotides. a) Eight nucleotides, including four natural nucleotides, A, T, C, and G, and four artificial nucleotides, P, Z, S, and B. A-T, C-G, P-Z, and S-B are orthogonal pairs. b) Three types of source files, text, picture, and video, encoded by the RaptorQ-Arithmetic-Base64-RS (RABR) system using eight nucleotides and their average coding efficiencies. c) Schematics of the encoding process of the RABR and RALR systems. The RABR system includes the systematic RaptorQ encoder, Arithmetic encoder, improved Base64 encoder and Reed–Solomon (RS) encoder, while improved Base64 encoder is replaced by the improved Lempel–Ziv–Welch (LZW) encoder in the RALR system.

by improved Base64 encoder or improved LZW encoder. iv) Errors inevitably happen within each DNA sequence during the processes of DNA synthesis and DNA sequencing. Introducing redundancy for error corrections within each DNA sequence is achieved by RS encoder. Therefore, the RABR and RALR systems could accommodate the characteristics of DNA storage techniques and are suitable for encoding DNA information.

The RABR and RALR systems first perform RaptorQ encoding with flexible control over redundancy. Arithmetic encoding instead of Huffman encoding is then used to compress the data. The codewords obtained by the ternary Huffman encoding are mapped to nucleotides one by one through the ternary rotation code.[29] In contrast, a unique arithmetic code can be generated for a specific sequence of length m, without

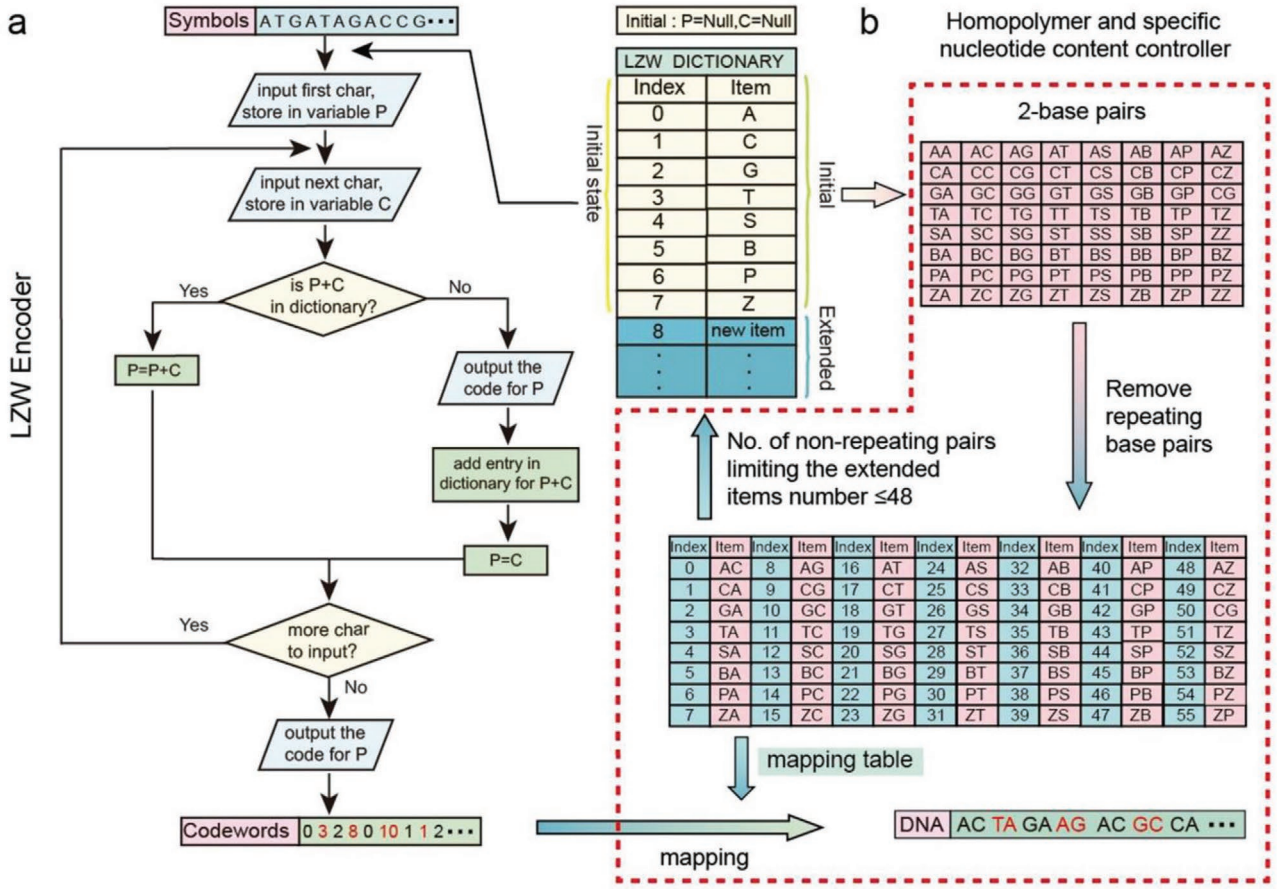

Figure 3. Schematics of improved LZW encoder for eliminating homopolymers and controlling the content of specific nucleotides. a) Flow diagram of improved LZW encoder. b) Mapping between LZW dictionary and mapping table. The LZW dictionary contains 56 different symbols or symbol sequences first encountered in the DNA sequence, each of which is designated with a mapping number. The combination of two different nucleotides creates 56 nonrepeating base pairs, each of which is also designated with a mapping number. The DNA sequence is mapped into a sequence of mapping numbers using the LZW dictionary, which is then mapped into a sequence of non-repeating base pairs using the mapping table.

the need for generating codewords for all possible sequences of length m. In addition, the coding efficiency of arithmetic compression can reach the level of source entropy in binary sources with highly inconsistent symbol probabilities.

The RABR and RALR systems adopt two different strategies to control homopolymers and GC contents to reduce the error rates in DNA sequencing. Imbalanced nucleotide contents may result in inapposite melting temperature and thus sequencing bias,[33] which will increase the error rates in DNA sequencing. Generally, each type of nucleotide with equal probability in DNA sequences is preferred. Therefore, the RABR system adopts Base64 encoding, code reshaping and balancing, and data mapping to reduce homopolymers and control GC contents[34] (Figure S3, Supporting Information). Initial binary data are first converted into Base64 codes,[35] which are then reshaped and converted into two groups, the binary codes and the balance codes. The binary codes and the balance codes are then mapped into a DNA sequence according to a customized mapping rule, in which "C" or "G" be mapped only when "1" appears in the balance codes. Since "1" appears in the balance codes with a 50% probability, GC contents in the DNA sequences are controlled at ≈50%. Homopolymers, i.e., continuous base repeat, such as CCC, are also reduced by the algorithm. The improved Base64 encoder for four nucleotides could be tuned for six and eight nucleotides. Different from the RABR system, the RALR system adopts improved LZW encoder, a dictionary compression algorithm, to eliminate homopolymers and control GC contents, as shown in **Figure 3**, Figure S4 (Supporting Information), and Supporting Information 4. The improved LZW encoder first creates an LZW dictionary, which contains symbols of A, C, G, T, S, B, P, and Z and another 48 different symbol sequences first encountered in the DNA sequences. Each of the 56 symbols and symbol sequences in the LZW dictionary is designated with a mapping number. Meanwhile, combinations of two different nucleotides using A, C, G, T, S, B, P, and Z create 56 non-repeating base pairs, each of which is randomly designated with a mapping number. In the improved LZW encoder, DNA sequences are first mapped into sequences of mapping numbers using the LZW dictionary and then mapped back into sequences of non-repeating base pairs using the mapping table, which ensures that there are no three-repeating nucleotides in the DNA sequences. Therefore, after LZW encoding, homopolymers are eliminated and each type of nucleotide relatively has an equal probability. Similarly,

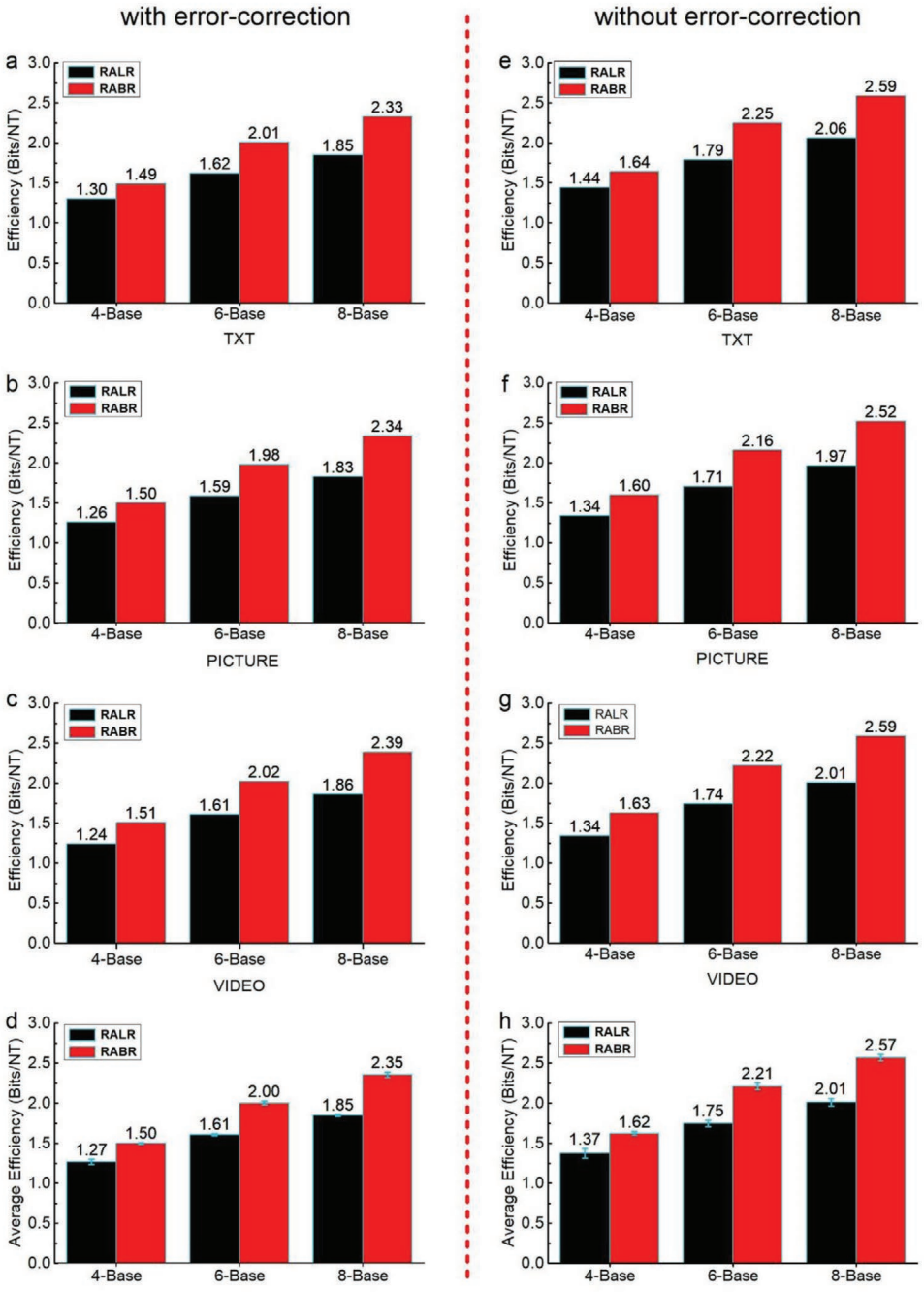

Figure 4. Coding efficiencies without arithmetic compression of text, pictures and videos encoded by the RALR or RABR system using four, six, or eight nucleotides with or without error correction. Coding efficiencies without arithmetic compression of a) text, b) picture, c) video, and d) their average by the RALR or RABR system with error correction. Coding efficiencies without arithmetic compression of e) text, f) picture, g) video, and h) their average by the RALR or RABR system without error correction.

the improved LZW encoder for eight nucleotides could slightly be adjusted for four and six nucleotides.

Since information is encoded in multiple DNA sequences, each DNA sequence requires a reasonably designed index, which allows the reorganization of source data stored in DNA sequences for the successful decoding of DNA encoded data. Composite indices, which are free of homopolymers, are generated by an index generation algorithm according to the number of source blocks and symbols. The number of nucleotides required for indexing is based on the formula $N \leq d^m - q$, where N, d, m, and q denote the number of DNA sequences, the number of nucleotide types, the number of nucleotides, and the number of combinations of m nucleotides containing homopolymers, respectively. All possible combinations of m nucleotides except those containing homopolymers are designated with a mapping number, which is added in the front of

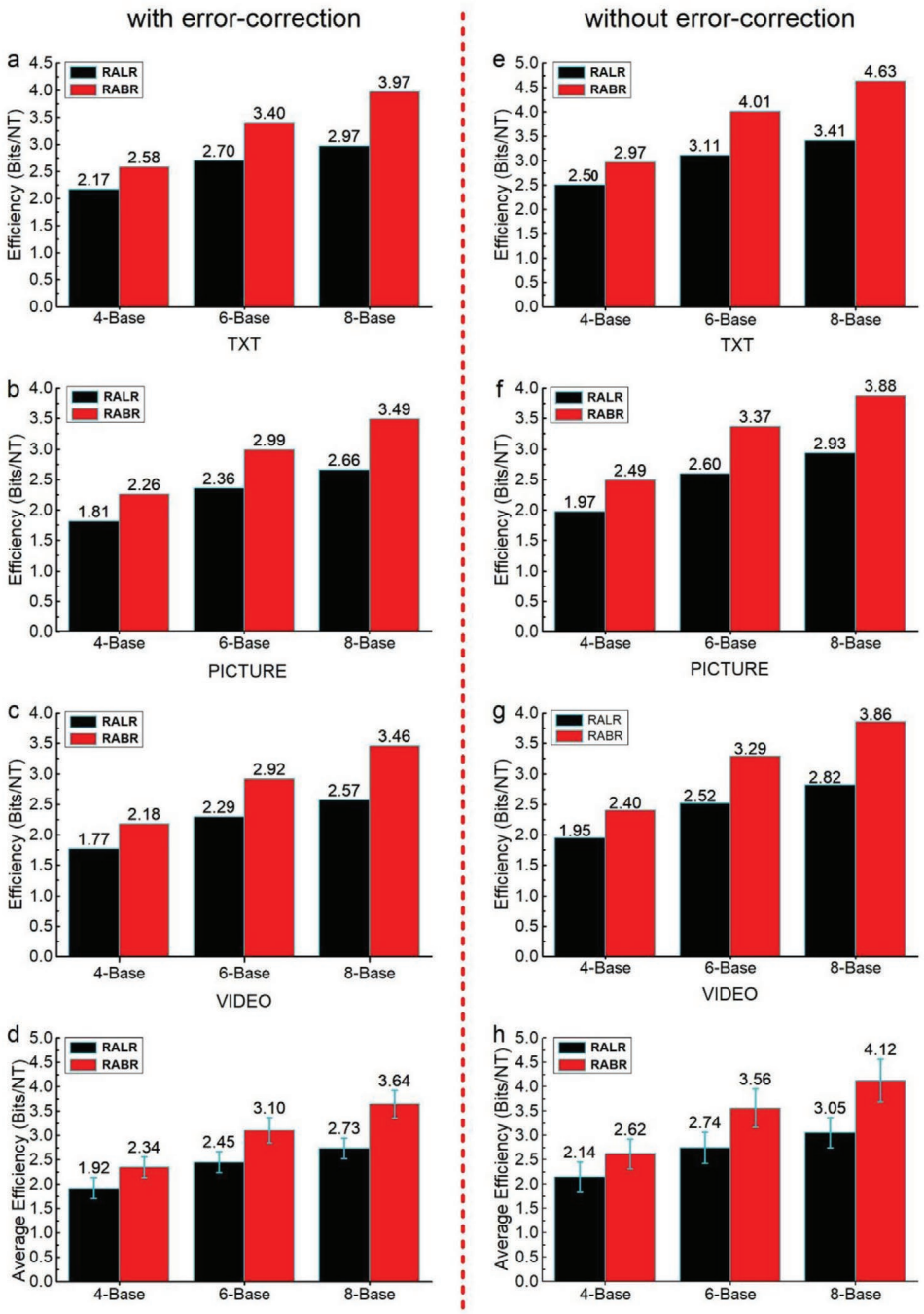

Figure 5. Coding efficiencies with arithmetic compression of text, pictures and videos encoded by the RALR or RABR system using four, six, or eight nucleotides with or without error correction. Coding efficiencies with arithmetic compression of a) text, b) picture, c) video, and d) their average by the RALR or RABR system with error correction. Coding efficiencies with arithmetic compression of e) text, f) picture, g) video, and h) their average by the RALR or RABR system without error correction.

each DNA sequence to number each DNA sequence, as demonstrated in Figure S5 (Supporting Information).

After indexing, each DNA sequence is encoded by the RS encoder, which corrects errors of substitution in the DNA sequence. Generating polynomials are first calculated in the Galois field; RS codeword polynomials are then obtained by multiplying the information polynomials with the generating polynomials; finally, the corresponding RS encoding symbols are obtained from the coefficients of RS codeword polynomials. After RS encoding, binary symbols are mapped to DNA symbols for DNA synthesis. Because increasing the error correction ability, i.e., the reliability of encoding, will inevitably increase the redundancy and thus reduce the coding efficiency, a balance between reliability and coding efficiency needs to

be achieved by analyzing the characteristics of DNA synthesis and sequencing. The error rate in DNA sequencing is roughly 1%.[9,18] The reliability of the encoding system is achieved by the RaptorQ encoder, which recovers data losses caused by DNA sequence losses, and RS encoder, which recovers errors caused by nucleotide substitution errors within each DNA sequence. The detailed algorithm of RS is described in Supporting Information 6.

To test the RABR and RALR encoding systems, DNA data storage using four, six, and eight nucleotides is encoded via python language. The coding efficiencies without arithmetic compression of text, picture, and video encoded by the RABR or RALR systems using four, six, or eight nucleotides with or without error correction are shown in **Figure 4**, while those with arithmetic compression are shown in **Figure 5**. The strong reliability of the RABR and RALR systems in processing large data is systematically shown in Tables S1–S4 (Supporting Information). The length of redundancy for DNA sequence loss and substitution errors within each DNA sequence could be tailored according to the error rate generated by the DNA data storage channel. The two encoding systems have achieved a relatively high coding efficiency, which could greatly reduce the cost of actual DNA data storage. With error correction and without arithmetic compression, the average coding efficiencies by the RALR system using four, six, and eight nucleotides reach 1.27, 1.61, and 1.85 bits per nucleotide, respectively. While the average coding efficiencies by the RABR system using four, six, and eight nucleotides are up to 1.50, 2.00, and 2.35 bits per nucleotide, respectively. With error-correction and arithmetic compression, the average coding efficiencies by the RALR system using four, six, and eight nucleotides reach 1.92, 2.45, and 2.73 bits per nucleotide, respectively. While those by the RABR system using four, six, and eight nucleotides reach 2.34, 3.10, and 3.64 bits per nucleotide, respectively. The RABR system has a slightly higher coding efficiency than that of the RALR system. This is because in the process of LZW encoding in the RALR system mapping of one single nucleotide, i.e., A, C, G, T, S, B, P, and Z, to two different nucleotides may happen in large probability, while the other 48 different symbol sequences first encountered in the DNA sequences may not have a high probability. However, the RALR system can completely eliminate homopolymers, while the RABR system can only reduce the probability of homopolymers. Both the RABR and RALR systems can qualitatively control the GC contents to a suitable proportion. To show the feasibility of the RABR and RALR systems, encoding of the poem "THE VALUE OF TIME" into DNA sequences was conducted (Figures S6 and S7, Supporting Information). Then successful decoding from perturbed DNA sequences with DNA sequence loss and substitution error by the RABR and RALR systems was realized and the correct information back to "THE VALUE OF TIME" was achieved.

4. Conclusion

Two encoding systems (RABR and RALR) with high reliability and coding efficiency, which contain a cascade of RaptorQ fountain encoder, Arithmetic encoder, improved Base64 encoder or improved LZW encoder, and RS correction encoder, are designed according to the characteristics of the DNA storage channel. By extending the bases to eight nucleotides, including four natural nucleotides and four artificial nucleotides, the maximum coding efficiency is increased to 3 bits per nucleotide. The reliability of the systems is achieved through RaptorQ encoding and RS encoding, which recover DNA sequence losses and nucleotide substitution errors within each DNA sequence, respectively. Improved Base64 encoding or LZW encoding is implemented in the system to control homopolymers and specific nucleotide content and reduce DNA sequencing errors. Relatively high coding efficiency is achieved by optimizing the balance between reliability and coding efficiency. The coding efficiency could further be increased by arithmetic compression. The RABR and RALR encoding systems are applicable for four, six, and eight nucleotides and could encode different types of source files, which offer a meaningful reference for the application of artificial nucleotides in DNA data storage.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The research was supported by the National Key R&D Program of China (2018YFA0902600, 2021YFF1200300, and 2020YFA0712102), the National Natural Science Foundation of China (Grant No. 21 877 104, 21 834 007, 21 907 088, 21 878 258, 22 020 102 003, and 22 125 701), the Youth Innovation Promotion Association of CAS (Grant No. 2 020 228), and the Zhejiang Provincial Natural Science Foundation of China (Grant No. Y20B060027).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article.

Keywords

artificial nucleotides, data storage, DNA, encoding systems, high-density data storage

Received: October 25, 2021
Revised: January 5, 2022
Published online: February 10, 2022

[1] X. Lu, T. Ellis, *Natl. Sci. Rev.* **2021**, *8*, nwab086.
[2] M. Gu, Q. Zhang, S. Lamon, *Nat. Rev. Mater.* **2016**, *1*, 16070.
[3] L. Organick, B. H. Nguyen, R. McAmis, W. D. Chen, A. X. Kohll, S. D. Ang, R. N. Grass, L. Ceze, K. Strauss, *Small Methods* **2021**, *5*, 2001094.

[4] Z. Li, C. Wang, J. Li, J. Zhang, C. Fan, I. Willner, H. Tian, *CCS Chem.* **2020**, *2*, 707.
[5] F. De Carli, N. Menezes, W. Berrabah, V. Barbe, A. Genovesio, O. Hyrien, *Small Methods* **2018**, *2*, 1800146.
[6] M. Campbell, *Computer* **2020**, *53*, 63.
[7] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, W. J. Stark, *Angew. Chem., Int. Ed.* **2015**, *54*, 2552.
[8] J. P. L. Cox, *Trends Biotechnol.* **2001**, *19*, 247.
[9] L. Ceze, J. Nivala, K. Strauss, *Nat. Rev. Genet.* **2019**, *20*, 456.
[10] J. Sun, B. Li, F. Wang, J. Feng, C. Ma, K. Liu, H. Zhang, *CCS Chem.* **2021**, *3*, 1669.
[11] J. Koch, S. Gantenbein, K. Masania, W. J. Stark, Y. Erlich, R. N. Grass, *Nat. Biotechnol.* **2020**, *38*, 39.
[12] Y. Hao, Q. Li, C. Fan, F. Wang, *Small Struct* **2020**, *2*, 2000046.
[13] R. Heckel, I. Shomorony, K. Ramchandran, D. N. C. Tse, in 2017 IEEE Int. Symp. Inform. Theory **2017**, 3130.
[14] D. Yiming, F. Sun, Z. Ping, Q. Ouyang, L. Qian, *Natl. Sci. Rev.* **2020**, *7*, 1092.
[15] K. Matange, J. M. Tuck, A. J. Keung, *Nat. Commun.* **2021**, *12*, 1358.
[16] D. Sharma, R. Kumar, M. Gupta, T. Saxena, *IET Nanobiotechnol.* **2020**, *14*, 635.
[17] T. P. Niedringhaus, D. Milanova, M. B. Kerby, M. P. Snyder, A. E. Barron, *Anal. Chem.* **2011**, *83*, 4327.
[18] R. Heckel, G. Mikutis, R. N. Grass, *Sci. Rep.* **2019**, *9*, 9663.
[19] J. Li, Y. Sun, Y. Liang, J. Ma, B. Li, C. Ma, R. E. Tanzi, H. Zhang, K. Liu, C. Zhang, *CCS Chem.* **2021**, *3*, 1830.
[20] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, C. N. Takahashi, S. Newman, H.-Y. Parker, C. Rashtchian, K. Stewart, G. Gupta, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze, K. Strauss, *Nat. Biotechnol.* **2018**, *36*, 242.
[21] G. Church, Y. Gao, S. Kosuri, *Science* **2012**, *337*, 1628.
[22] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, E. Birney, *Nature* **2013**, *494*, 77.
[23] R. Gallager, in IEEE Trans. Inf. Theory **1978**, *24*, 668.
[24] M. Blawat, K. Gaedke, I. Huetter, X.-M. Chen, B. Turczyk, S. Inverso, B. Pruitt, G. Church, *Procedia Comput. Sci.* **2016**, *80*, 1011.
[25] L. Anavy, I. Vaknin, O. Atar, R. Amit, Z. Yakhini, *Nat. Biotechnol.* **2019**, *37*, 1229.
[26] L. Meiser, P. Antkowiak, J. Koch, W. Chen, X. Kohll, W. Stark, R. Heckel, R. Grass, *Nat. Protoc.* **2019**, *15*, 86.
[27] M. Luby, A. Shokrollahi, M. Watson, T. Stockhammer, L. Minder, RaptorQ Forward Error Correction Scheme for Object Delivery. *Internet Engineering Task Force* **2011**.
[28] Y. Erlich, D. J. S. Zielinski, *Science* **2017**, *355*, 950.
[29] J. Bornholt, R. Lopez, D. Carmean, L. Ceze, G. Seelig, K. Strauss, *EEE Micro* **2017**, *37*, 98.
[30] S. Hoshika, N. A. Leal, M. J. Kim, M. S. Kim, N. B. Karalkar, H. J. Kim, A. M. Bates, N. E. Watkins, H. A. Santalucia, A. J. J. S. Meyer, *Science* **2019**, *363*, 884.
[31] W. Lu, R. Hu, X. Tong, D. Yu, Q. Zhao, *Small Struct.* **2020**, *1*, 2000003.
[32] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, D. B. Jaffe, *Genome Biol.* **2013**, *14*, R51.
[33] K. Chen, J. Zhu, F. Bošković, U. F. Keyser, *Nano Lett.* **2020**, *20*, 3754.
[34] Y. Zhang, L. Kong, F. Wang, B. Li, C. Ma, D. Chen, K. Liu, C. Fan, H. Zhang, *Nano Today* **2020**, *33*, 100871.
[35] S. M. H. T. Yazdi, R. Gabrys, O. Milenkovic, *Sci. Rep.* **2017**, *7*, 5011.