DNA Computing-Based Multi-Source Data Storage Model in Digital Twins

JINXIA WANG, School of Art and Design, Shaanxi Fashion Engineering University, Xian, China
RUI CHEN, Xi'an University of Posts & Telecommunications, Xi'an, China
ZHIHAN LV, Department of Game Design, Faculty of Arts, Uppsala University, Sweden

The work aims to study the application of **Deoxyribonucleic Acid (DNA)** multi-source data storage in **Digital Twins (DT)**. Through the investigation of the research status of DT and DNA computing, the work puts forward the concept of DNA multi-source data storage for DT. Raptor code is improved from the design direction of degree distribution function, and six degree function distribution schemes are proposed in turn in the process of describing the research method. Additionally, a quaternary dynamic Huffman coding method is applied in DNA data storage, combined with the improved concatenated code as the error correction code. Considering the content of **cytosine deoxynucleotide (C)** and **guanine deoxynucleotide Guanine (G)** and the distribution of homopolymer in DNA storage, the work proposes and verifies an improved concatenated code algorithm **Deoxyribonucleic Acid-Improved Concatenated code (DNA-ICC)**. The results show that while the **Signal-to-Noise Ratio (SNR)** increases, the **Bit Error Rate (BER)** decreases gradually and the trend is similar. But the anti-interference ability of the degree distribution function optimized by the probability transfer method is better. The BER of DNA-ICC scheme decreases with the decrease of error probability, which is stronger than other error correction codes. Compared with the original concatenated code, it saves at least 1.65 s, and has a good control effect on homopolymer. When the size of homopolymer exceeds 4 nt, the probability of homopolymer is only 0.44%. The proposed Quaternary dynamic Huffman code and concatenated error correction code have excellent performance.

CCS Concepts: • **Hardware** → *Communication hardware, interfaces and storage*; *External storage*;

Additional Key Words and Phrases: Digital Twins, DNA computing, multi-source data storage, Huffman coding, error correction code

ACM Reference format:
Jinxia Wang, Rui Chen, and Zhihan Lv. 2023. DNA Computing-Based Multi-Source Data Storage Model in Digital Twins. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 3s, Article 127 (February 2023), 16 pages.
https://doi.org/10.1145/3561823

1 INTRODUCTION

In recent years, with the rapid development of Cloud Computing, Big Data, **Internet of Things (IoT), Artificial Intelligence (AI)**, social networking and other information technology fields and the digital transformation of traditional industries, human beings have generated more and more

Authors' addresses: J. Wang, School of Art and Design, Shaanxi Fashion Engineering University, No. 1, Tongwen Road, Fengxi New Town, Xixian New Area, Xi'an, 712046, China; email: 1754354519@qq.com; R. Chen, Xi'an University of Posts & Telecommunications, No. 563, Chang'an South Road, Xi'an, 710061, China; email: chenrui@xupt.edu.cn; Z. Lv (corresponding author), Department of Game Design, Faculty of Arts, Uppsala University, Strandgatan 1b (Residenset), Visby, 62167, Sweden; email: zhihan.lyu@speldesign.uu.se.

1551-6857/2023/02-ART127
https://doi.org/10.1145/3561823

information. This requires not only the storage capacity, but also the scalability of storage time and space, fault tolerance, and error correction of stored procedures. With the arrival of the Big Data era, the imagination of the sharp increase in data capacity has led to the fact that the traditional data storage methods can no longer meet the demand, and the demand for new storage media has been paid more and more attention to [1]. **Deoxyribonucleic Acid (DNA)** data storage is considered as a medium suitable for long-term data storage due to its ultra-high density and sufficiently stable storage performance [2–4]. DNA data storage can reach a storage density of 1018byte/mm^3, which is almost 106 times higher than the current highest storage density media. It has very good anti-interference ability against external high temperature and vibration, and can be preserved for a hundred years [5]. The popularity of **Digital Twins (DT)** has been rising in recent years, which has attracted much attention from the industry. It can make full use of physical model, sensor update, operation history and other data, integrate multi-disciplinary, multi physical quantity, multi-scale, and multi probability simulation process, and complete mapping in virtual space, thus reflecting the whole life cycle process of corresponding physical equipment [6–8]. One of the key features of DT is multi-source heterogeneous data fusion. The visual decision-making system also pays attention to the integration and comprehensive application of multi-source heterogeneous data [9]. Tremendous basic data will be generated during the actual operation of various industries, including various map element data, monitoring video data, real-time message data, urban tilt photography data, sensor data, business system data, and various database data. The visual decision-making system based on DT data fusion and DNA data storage can fully integrate the massive data between different departments, industries, systems, and data formats, and provide comprehensive data support for the perception and judgment of operation situation in various fields [10–12].

At present, the hard disk and other storage systems widely used by people have inherent shortcomings. For example, the storage life of hard disk and flash memory is only a few decades at most. The storage equipment is non-degradable and pollutes the environment. It is urgent to develop a new generation of alternative storage technology. DNA computing is a non-traditional computing technology based on DNA and enzymes, and depends on the principles of biochemistry and molecular biology [13]. DNA strands can be used to encode and store information. Researchers have designed a variety of mapping strategies, such as differential mapping and constraint mapping, to meet these biochemical constraints in the process of DNA computing. However, most mapping strategies have limited mapping potential at each nucleotide storage site [14–16]. In addition, error correcting codes, such as single parity check codes and duplicate codes, will be used in the storage of DNA data. Wang et al. (2021) [17] studied and discussed the information needs of China's Online Health Community for Corona Virus Disease 2019. These error correction methods often have the problems of high complexity and decoding failure. Therefore, a more stable DNA data storage method is needed to solve the existing problems [18]. Any new technology requires a shift from theory to practical application, as is the case with DT. The concept of DT originates from the industrial manufacturing field. Driven by 5G communication, IoT, Cloud Computing, Big Data, AI and other new generation information technologies, the concept of DT is gradually extended to more industry spaces [19–21].

With the gradual deepening of people's understanding of DNA and the rapid development and maturity of DNA storage related technologies, researchers gradually consider DNA as a new information storage medium. Qian et al. [22] pointed out that only a few studies had solved the relationship between data, and most of them were conducted outside the operating environment. Data visual decision-making can be quickly pushed into the application of DT industries to help industry managers improve their intelligent decision-making ability and efficiency. DNA storage not only has large storage capacity, high density, long storage life, but also has higher security, energy conservation, and environmental protection. As a cross fusion technology of information

technology and biotechnology, DNA information storage technology plays an important role in saving storage energy and promoting the development of massive data storage. However, the research on DNA information storage is still in its infancy in China, and more energy, human and material resources need to be invested. From the perspective of long-term investment, many manufacturers believe that this technology has high value, and this technology is likely to become a breakthrough in the search for new storage media in the future.

Through literature research and algorithm verification, this work studies the problem of DNA data storage for DT. The innovation lies in the improvement of Raptor code, and six degree distribution function distribution schemes are proposed. Based on the original Huffman code, a quaternary dynamic Huffman code is proposed for the encoding and decoding of DNA data storage. Considering the content of cytosine deoxynucleotide C and guanine deoxynucleotide G and the distribution of homopolymer in DNA storage, an **improved concatenated code (ICC)** is proposed to be used as error correction code, and its performance is verified to be excellent.

2 RESEARCH STATUS OF DT AND DNA COMPUTING

The DNA computer used in the DNA computing process has many characteristics, such as small size, large storage capacity, fast operation, low energy consumption, and parallelism, and has advantages in data storage [23–25]. The visual data system with higher quality and requirements can be realized via the MHD fusion technology in DT. In the development of DNA computing, scholars were faced with complex problems, such as detecting the operation results of logic gates and judging whether the Logic Gates were successfully constructed. Later, scholars solved these problems by labeling various Molecule Radicals on the Thymine (T) Base. With the development of DNA computing, Logic Gates have gradually evolved into Logic Circuits, and many experts and scholars have given answers in this area.

Most research on DT only focuses on existing explicit frameworks and architectures, which face the challenge of supporting different levels of integration through agile processes [26]. Aheleroff et al. [27] conducted a study to determine the appropriate Industry 4.0 technology and the overall reference architecture model to achieve the most challenging DT application. With the intensification of market competition, the process of product development is accelerated, which requires rapid product innovation and efficient collaboration between design and manufacturing. However, there are still islands of information that hinder the integration of product lifecycle processes. Bionics and DT have been combined as potential solutions to address this problem. Some scholars proposed the concept, framework, and characteristics of DT bionics and expounded on the co-evolution mechanism of product twin, including virtual and physical product and production twin. Li et al. [28] put forward a symbiotic and co-evolutionary mechanism to integrate product development and manufacturing. They concluded that integrating bionics and DT could accelerate the innovation and development of new products and help realize the effective management of production construction.

In addition to data security, long search time for data owners, long data access time, and high system leakage rate may occur when they access data from the cloud environment. Given all the above problems, Namasudra et al. [29] introduced a fast and secure data access control model based on DNA for cloud environment. In the model, cloud service providers needed to maintain a fast and efficient data access table. The authors used a long cipher or key based on 1,024 bits of DNA to encrypt a user's confidential or personal data. They finally verified the effectiveness of the proposed model compared with existing models through experimental results and theoretical analysis. Adithya and Santhi [30] provided a color code encryption strategy of DNA computing to protect data from eavesdroppers. DT is often defined as real-world products, systems, existence, communities, and even cities, using virtual copies of data from their physical counterparts and their

environments that are constantly updated. It connects virtual cyberspace with physical entities, so it is considered to be the pillar of industry 4.0 and the innovation in the future. It can be said that DT is created and used in the whole life cycle of the entity it copies, from cradle to grave. Jiang et al. [31] focused on the current situation of DT and its application in industry under the background of intelligent manufacturing, especially from the perspective of plant wide optimization. In this context, the main functions of DT are discussed, such as mirroring, ghosting, and threading.

To sum up, DT and DNA computing have achieved certain research results, and the relevant research conclusions have their own advantages and disadvantages. However, the research on DNA data storage mostly focuses on the data storage security, and pays less attention to the efficiency of data storage itself. This work optimizes and improves different encoding and decoding schemes, which can provide new ideas and research directions for the future research of DNA multi-source data storage.

3 DNA MULTI-SOURCE DATA STORAGE IN DT WITH HUFFMAN ENCODING AND CASCADE ECCS

3.1 DNA Multi-source Data Storage for DT

DNA data storage unit is four Deoxynucleotide, namely Adenine (A), Thymidine (T), Cytosine (C), and Guanine (G), also known as Bases. These four Nucleotide Groups can be arranged in different ways to form Oligonucleotides, i.e., DNA Strands. DNA data storage consists of three basic structures: data writing and reading [32, 33, 34]. The data writing part includes encoding, mapping, and composition. The data to be stored is input in Binary and encoded by Source Compression and Channel Error Correction. Then, some mapping rules transform the Binary Sequence into the Sequence consisting of four Nucleotide Bases. Finally, DNA chains are synthesized by biochemical methods and stored independently in special containers. Current DNA synthesis techniques can synthesize DNA strands of 200 to 1,000 bases in length. When reading the data, the Polymerase Chain Reaction [35, 36, 37] amplification technology should be used to copy the data of Oligonucleotides in the storage pool. A typical sequencing method relies on the characteristic that Fluorescent Nucleotides emit different colors. Precisely, the DNA sequence represented by the oligonucleotides can be read out by detecting the colors. Figure 1 illustrates the specific DNA data storage process.

In Figure 1, the biological read-writer technology involved in DNA data storage include DNA synthesis, PCR amplification, and DNA sequencing. In view of the biological activities and the cell's exclusion, DNA synthesis is usually performed artificially in vitro. The synthesis process is divided into three stages. First, the Base Sequence is divided into several short strands because after the information is encoded, the Base Sequence is longer. Then, address Bits are added to the short chain after segmentation, convenient for quick search, location, and splicing when the subsequent file is read. Finally, it is necessary to add Macromolecular Primers with Nucleotide Sequences at both ends of the DNA chain and then preserve them to facilitate the splicing of short DNA chains and realize the function of data access.

3.2 Research on the DNA Information Storage Method based on Raptor Code

A key feature of Fountain Code is recovering information symbols using very small decoding with a high probability of receiving overhead. Fountain Code can be decoded successfully after receiving a certain number of code symbols [38]. However, **LT (Luby transform)** Code can be decoded 100% successfully only after all original signals are recovered during decoding, and the decoding complexity is not linear. The precoding link is added to the Raptor Code to improve coding performance and ensure the same synthesis cost, effectively resolving the contradiction between the complexity of coding and decoding and transmission efficiency. The precoding part

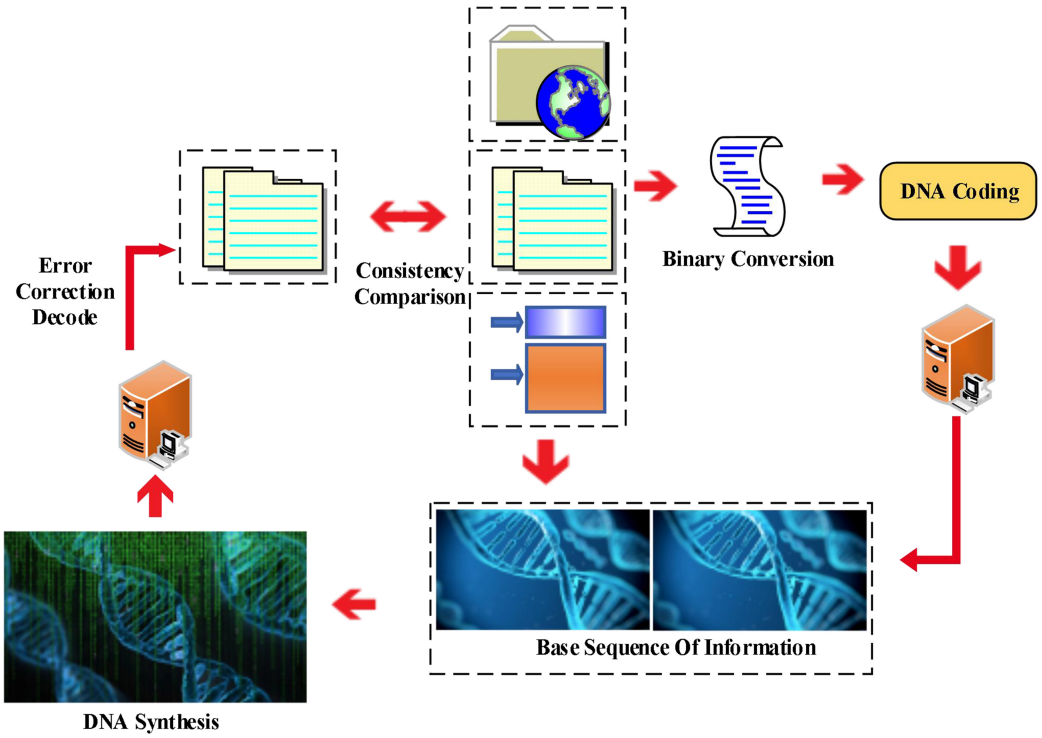

Fig. 1. Implementation process of DNA information storage.

primarily uses low density parity check for **LDPC (Low Density Parity Check Code)**. Raptor Code is cascaded LDPC and LT Code. Assuming that there is a linear block code with a code length of N, in which the number of information symbols is K and the number of check codes is M, then the matrix generated by this block code is defined as $G_{k\times N}$. In other words, the information symbol matrix $U_{k\times 1}$ can be mapped to the block code space through the generation matrix, and the block code matrix C and check bits M meet Equations (1) and (2):

$$M = N - k \tag{1}$$

$$C = U \times G \tag{2}$$

For irregular LDPC, the degree value of each node is determined by the Degree Distribution function. Equation (3) describes the Degree Distribution function.

$$\lambda(x) = \sum_{j=2}^{d_v} \lambda_j \, x^{j-1} \tag{3}$$

In Equation (3), λ_j represents the ratio of the number of edges s_j owned by the variable node with degree value j to the total number of edges z of the bidirectional graph, which can be expressed as Equation (4).

$$\lambda_j = \frac{s_j}{z} \tag{4}$$

Besides, d_v refers to the maximum degree value of the variable node, satisfying:

$$\sum_{i=2}^{d_v} \lambda_i = 1 \tag{5}$$

Then, the Degree Distribution function of the verification node is expressed as Equation (6).

$$\rho(x) = \sum_{i=2}^{d_c} \rho_i \, x^{i-1} \tag{6}$$

In Equation (6), ρ_i denotes the ratio of the number of edges b_j owned by the check node whose degree value is i to the total number of edges z in the bidirectional graph, as shown in Equation (7).

$$\rho_i = \frac{b_i}{z} \tag{7}$$

d_c in Equation (6) signifies the maximum degree value of the verification node, and it satisfies Equation (8).

$$\sum_{i=2}^{d_c} \rho_i = 1 \tag{8}$$

The core of LDPC is to determine the check matrix H to obtain the generation matrix G and then encode the information symbol. The length of the information symbol determines the dimension of the check matrix H. Assuming that the length of the information symbol waiting to be encoded is n, then there is a $m \times n$ dimensional check matrix. The Gaussian elimination method is used to transform this check matrix into the standard form, that is:

$$H^{\cdot} = [p| \, I_m] \tag{9}$$

where p stands for a $m \times (n - m)$ dimensional check matrix, and I_m denotes a m-dimensional identity matrix. After standardization by the Gaussian elimination method, the identity matrix can be written as Equation (10).

$$p_i = \sum_{j=1}^{n-m} H'_{i,j} \, s_j + \sum_{j=1}^{i-1} H'_{i,j+n-m} \, p_j \tag{10}$$

The **Belief Propagation (BP)** decoding algorithm is adopted to decode LDPC. Firstly, variable nodes are initialized, and symbols are assigned according to the acceptance conditions in Equation (11).

$$p_i = \begin{cases} +\infty & y_i = 0 \\ 0 & y_i = E \\ -\infty & y_i = 1 \end{cases} \tag{11}$$

In Equation (11), E represents the variable node to be deleted. The Exclusive OR operation is performed to delete the associated edges between all the remaining n variable nodes that have not been deleted and the check nodes connected to these nodes, as presented in Equation (12).

$$\phi_{mn} = 2 \tanh^{-1} \left(\prod_{n' \in N(m)|n} \tanh \left(\frac{\varphi_{mn'}}{2} \right) \right) \tag{12}$$

In Equation (12), $N(m)|n$ represents the number of check nodes connected to the remaining variable nodes. Assuming that the associated edge of a particular check node is 1, then the variable node connected to the node can be restored. The value of the check node is $N(m)$. After that, the associated edges connected to the restored variable node can be deleted, which can be written as:

$$\phi_{mn} = \phi_{n0} + \sum_{m' \in M(n)|m} \phi_{m'n} \tag{13}$$

Table 1. BP Decoding Algorithm

1	**Start** Initialize the variable node and assign the symbol as 0, E, 1 according to the reception:
2	$p_i = \begin{cases} +\infty & y_i = 0 \\ 0 & y_i = E \\ -\infty & y_i = 1 \end{cases}$
3	Initialize all verification nodes to 0.
4	For all remaining variable nodes, XOR and delete the associated edges between them at the same time:
5	$\phi_{mn} = 2\tanh^{-1}\left(\prod_{n' \in N(m)\mid n} \tanh\left(\frac{\varphi_{mn'}}{2}\right)\right)$
6	**If** The associated edge of a check node is 1
7	$\phi_{mn} = \phi_{n0} + \sum_{m' \in M(n)\mid m} \phi_{m'n}$
8	Deletes the associated edge connected to the variable node.
9	**Else if**
10	No check node with associated edge 1 was found during decoding
11	Decoding aborted.
12	**End**

$$\phi_n = \phi_{n0} + \sum_{m \in M(n)} \phi_{mn} \tag{14}$$

Finally, if no check node with an associated edge of 1 can be found, the decoding will be terminated. The whole decoding process is the process of eliminating associated edges. The Table 1 is the specific algorithm process.

Raptor Code is improved by designing the Degree Distribution function. The Degree Distribution function with good performance needs to guarantee the coverage of coded symbols and minimize the average degree value as much as possible to reduce the complexity of encoding and decoding. Therefore, the Distribution function satisfying these two points can be expressed as Equation (15).

$$\begin{cases} 1 - d - e^{-\mu'(d)(1+\varepsilon)} \geq \gamma\sqrt{\frac{1-d}{k}} \\ d \in [0, 1-\delta] \end{cases} \tag{15}$$

In Equation (15), ε represents the coding and decoding redundancy of Fountain Codes, γ refers to a positive real number, and $1-\delta$ indicates the decoding success rate. Besides, k signifies the number of information symbols, and $\mu'(d)$ is the derivative of $\mu(d)$. Then, there is:

$$\begin{cases} A \geq \gamma\sqrt{(1-d)\cdot k} \\ d \in [0, 1-\delta] \end{cases} \tag{16}$$

where A represents the decoded set received by the receiver. Because DNA data is stored during coding, long strands are more error-prone, costly, and technically complex than short strands. In addition, in the process of Raptor Code coding, if the symbol code is long, the complexity in the coding process will be increased, and the overall timeliness will be reduced. Therefore, the short code length is often used to obtain a DNA-Raptor data storage architecture with better performance.

Common Degree Distribution functions satisfying the above requirements can be expressed as:

$$\mu(d) = 0.00098d + 0.459\,d^2 + 0.211\,d^3 + 0.113\,d^4 + 0.1113\,d^{10} + 0.0799\,d^{11} + 0.0156\,d^{40} \quad (17)$$

$$\mu(d) = 0.007969d + 0.4935\,d^2 + 0.166\,d^3 + 0.073\,d^4 + 0.082\,d^5 + 0.056\,d^8 + 0.037\,d^9 + 0.055\,d^{19} + 0.025\,d^{65} + 0.0031\,d^{66} \quad (18)$$

Since the success rate of LT decoding is positively correlated with redundancy, A slowly increases after reaching the peak value until it is close to full decoding. Therefore, the improved Degree Distribution function can be written as Equation (19).

$$\tau(d) = \begin{cases} \frac{s}{kd} & d = 1, 2, \ldots, \frac{k}{s} \\ 0 & d > \frac{k}{s} \end{cases} \quad (19)$$

Then, the **Robust Solitary Wave Distribution (RSWD)** $\mu(d)$ is adopted as Scheme 1, as shown in Equation (20).

$$\mu_1(d) = \frac{\rho(d) + \tau(d)}{\sum_{d=1}^{k} \rho(d) + \tau(d)} \quad (20)$$

Scheme 2 is proposed based on setting the degree value with low probability in the Robust Solitary Wave Distribution (RSWD) to 0, as presented in Equation (21).

There are two kinds of probability distribution in the theory of soliton division: the ideal soliton distribution and Robust soliton distribution (RSWD).

$$\mu_2(d) = \begin{cases} \mu_2(d) = 0 & \mu(d) < \frac{1}{k} \\ \mu_2(d) = \mu(d) & \mu(d) \geq \frac{1}{k} \end{cases} \quad (21)$$

A new Degree Distribution function is obtained after normalization:

$$\mu_2(d) = \frac{\mu_2(d)}{\sum_{d=1}^{k} \mu_2(d)} \quad (22)$$

Assuming that the number of information symbols of short code length k = 16~1024, then:

$$\mu(d) \geq \frac{-\ln(1 - d - \gamma\sqrt{\frac{1-d}{k}})}{1 + \varepsilon} \quad (23)$$

Due to the structural characteristics of DNA data storage and the performance characteristics of Raptor Codes, the symbol code length is set as $k = 256$. Besides, degree values d_{33} and d_{44} are added, the probability of occurrence of d_{65} is transferred to d_1, and then the probability of occurrence of d_{66} is transferred to d_{33}, which is taken as Scheme 3. The probability of the occurrence of $d_{65}$is shifted to d_1, and the probability of occurrence of d_{66} is shifted to d_{44}, which is regarded as Scheme 4. The occurrence probability of d_{65} is transferred to d_1, which is taken as Scheme 5. The occurrence probability of d_{45} is shifted to d_2, and the occurrence probability of d_{66} is shifted to d_{33}, which is regarded as Scheme 6. The specific expressions are as follows:

$$\mu_3(d) = 0.033d + 0.492\,d^2 + 0.167\,d^3 + 0.072\,d^4 + 0.082\,d^5 + 0.056\,d^8 + 0.037\,d^9 + 0.0556\,d^{19} + 0.003\,d^{33} \quad (24)$$

$$\mu_4(d) = 0.033d + 0.492\,d^2 + 0.167\,d^3 + 0.072\,d^4 + 0.082\,d^5 + 0.056\,d^8 + 0.037\,d^9 + 0.0556\,d^{19} + 0.003\,d^{44} \quad (25)$$

$$\mu_5(d) = 0.033d + 0.492d^2 + 0.167d^3 + 0.072d^4 + 0.082d^5 + 0.056d^8 + 0.037d^9 + 0.0556d^{19} + 0.003d^{66} \tag{26}$$

$$\mu_6(d) = 0.029d + 0.503d^2 + 0.167d^3 + 0.072d^4 + 0.082d^5 + 0.056d^8 + 0.037d^9 + 0.0556d^{19} + 0.003d^{33} \tag{27}$$

3.3 DNA Information Storage based on Adaptive Huffman Coding and Concatenated ECCs

Huffman Coding is an Entropy Coding algorithm used for lossless data compression. Code words of different lengths are allocated according to the probability of the occurrence of coded characters. The higher the probability, the shorter the code words; the lower the probability, the longer the code words. In this way, the storage density can be improved after average processing. However, some disadvantages include the encoding method with poor timeliness, some non-generic fields, and low encoding efficiency due to the additional space storing the Huffman Tree. This paper uses the **Adaptive Huffman Coding (AHC)** in DNA data storage to solve these problems. AHC dynamically adjusts the Huffman Tree every time the encoder reads the symbol to be encoded and changes the corresponding weight and Huffman Tree every time a character is read. In this way, it can ensure that the current output symbol is only related to the currently encoded character and the character read and has no relationship with the character not read. The decoding process is similar to the encoding process. In view of the advantages of quaternary coding, this paper proposes a QAHC algorithm for DNA data storage, namely DNA-QAHC.

DNA multi-source data storage for DT can also be regarded as a process of sending and receiving information, in which different degrees of noise interference will cause errors in the data transmission process. Although such errors are rare, they have profound implications for data recovery. Therefore, error correction is a must. The most common errors in DNA information storage usually occur in the process of data deletion, data insertion, and replacement, collectively known as synchronization errors. Here, Concatenated Codes, Watermark Codes, and non-binary LDPC are combined for error correction. Figure 2 reveals the specific error correction process.

As Figure 2 suggests, the decoding process can be defined as **Hidden Markov Model (HMM)** by associating sparse sequences with LDPC codes.

Errors with value a are inserted between the first bit t0 and the time tj to be sent, and errors with value b are deleted. The range of drift value xj of point j is:

$$X = \{-x_{\max}, \ldots, -1, 0, 1, \ldots, x_{\max}\} \tag{28}$$

where xmax represents the maximum drift value. The transition probability Pa,b is defined as Equation (29) to reduce the complexity in the decoding process.

$$P_{a,b} = P\left(x_{j+1} = b | x_j = a\right) \tag{29}$$

The concatenated code thus obtained has high complexity, large computational data redundancy, and low accuracy. Therefore, this work proposes an ICC algorithm and applies it to the DNA information storage process. Meantime, considering the content of cytosine deoxynucleotide (C) and guanine deoxynucleotide (G) and homopolymer in DNA storage, it is labeled as DNA-ICC algorithm. In other words, the ICC algorithm is applied to the DNA information storage process. In the algorithm, after calculating the forward and backward probabilities at the boundary of each transmitted symbol, only the ones with larger values are returned as possible drift states and corresponding symbols for subsequent decoding, and the decoding path is limited to a small range in the grid. This operation can avoid the path with very low probability from entering the calculation. The key to improve the concatenated code error correction scheme is to ensure the reliability of channel transmission.

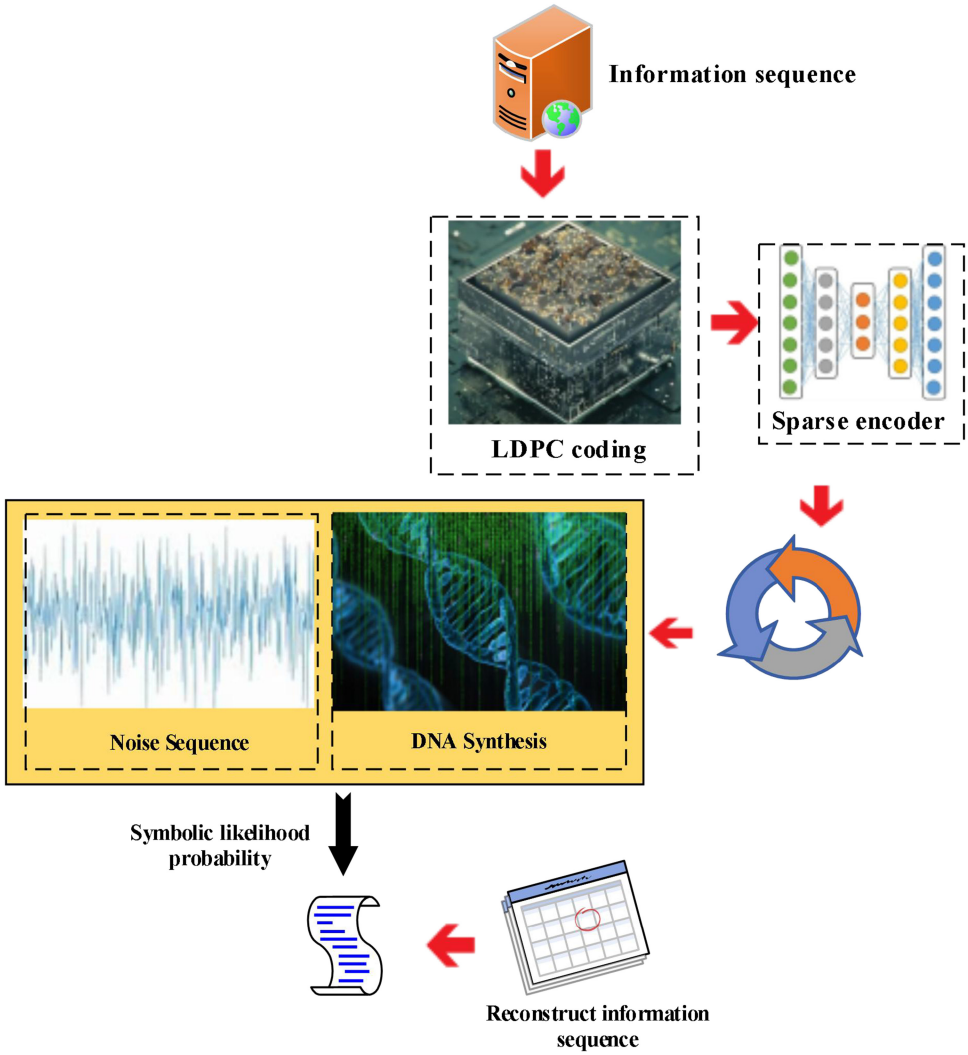

Fig. 2. The specific process of error correction by Concatenated Codes.

Additionally, the input sequence also has an impact on the error behavior. The guanine and cytosine base content and homopolymer are the factors that affect the high error probability in the process of DNA synthesis. Therefore, this work adopts the DNA-QAH algorithm to encode the original file and the concatenated code for error correction. In this way, after several different encodings, the scattered base sequence can be homogenized to a great extent, better guanine and cytosine base content and homopolymer can be obtained, and the error rate of data storage can be reduced.

3.4 Experimental Verification

Case analysis and performance comparison are conducted to verify the six Degree Distribution function distribution schemes. The length of the symbol code is set as 240, the verification symbol is 16 bits, and the code length is 256. Besides, the row weight of the verification matrix is 16,

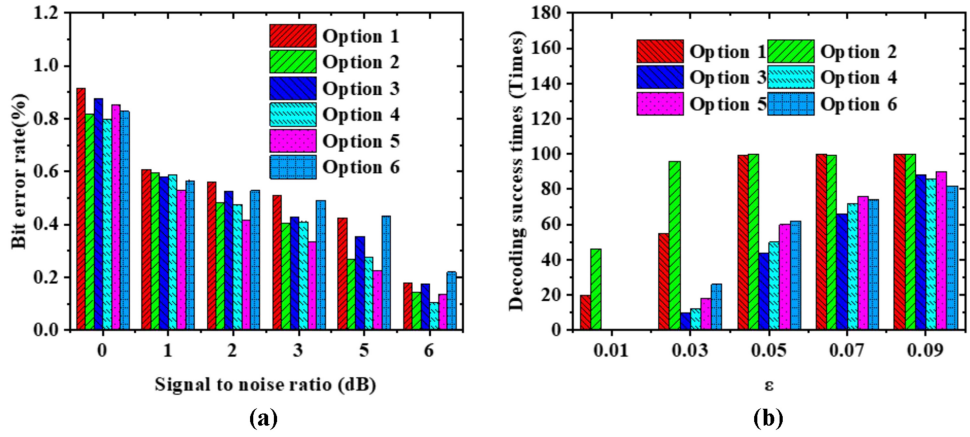

Fig. 3. Performance comparison of different Raptor Codes: (a) bit error rate under different SNR, (b) the number of successful decoding.

and the column weight is 1. To verify the performance of DNA-QAHC algorithm and DNA-ICC scheme, this paper selects different texts, images, audio, and video for case analysis and selects the Pseudo-Random Sequence as the Watermark Sequence. The operating system is 64-bit Windows 10, the processor is Intel Core i7-7500u, the memory size is 8GB, and the running platform is MatlabR2018a. The text, image, and audio selected by inputting file information are one of the data collected daily in the research process. The specific file information is shown in Table 2:

Table 2. Specific File Information

Encoding file type	Format	Memory (KB)	Note
Text 1	.txt	147.3	Chinese
Text 2	.txt	37.5	English
Image 1	.tif	11.8	Color image
Image 2	.tiff	25.4	Gray image
Audio 1	.wav	10.2	
Audio 2	.mp3	13.8	

4 RESULTS

4.1 Performance Comparison of Different DNA Information Storage Schemes based on Raptor Codes

Figure 3 illustrates the comparison of bit-error rates and decoding success times of the six schemes under different **Signal-to-Noise Ratios (SNRs)** and encoding and decoding redundancy ε of Fountain Codes.

In Figure 3(a), when the SNR increases, the error bit rate gradually decreases, and the six Raptor Code schemes have a similar downward trend. The SNR is set to 10dB in the case analysis to reduce the interference caused by SNR. Then, the redundancy of Fountain Code in Raptor Code is set to 0.01, 0.03, 0.05, 0.07, and 0.09. Figure 3(b) indicates that when decoding 100 times under the same SNR, Scheme 1 and Scheme 2 have a larger successful decoding rate than other schemes, which is close to 100% when the redundancy is greater than 0. 05. When the redundancy of Scheme 6 is

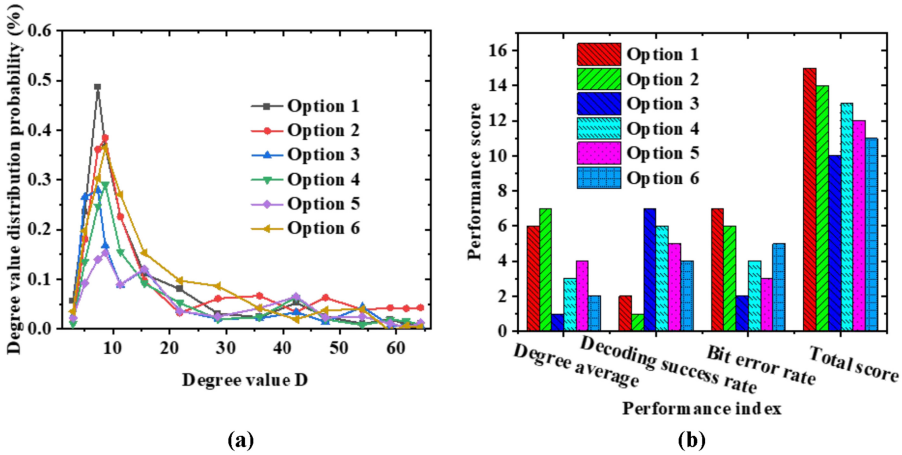

Fig. 4. Performance score distribution of different Raptor Code schemes: (a) degree distribution probability, (b) average degree, decoding success rate, and bit error rate performance score.

greater than 0. 09, the decoding rate is close to 90%. Therefore, these schemes have a small average degree, relatively low code complexity, and high decoding success rate.

The degree value distribution probability, average degree, decoding success rate, and bit error rate performance of different Raptor Code schemes are ranked in order of scoring, with the performance decreasing from 1 to 7, as shown in Figure 4.

Figure 4(a) shows that the degree value distribution probabilities of the six Raptor Code schemes are very similar. With the increase in degree value, the variation trend of distribution probability and the probability of degree value are very similar. Figure 4(b) indicates that a higher average degree can increase the complexity of the encoding and decoding but can guarantee coverage in the process of signal transmission. The scores suggest that the higher the score of degree average, the lower the corresponding score of the decoding success rate. This result proves that decoding the success rate will also increase with the increase of the average degree.

4.2 Performance Verification of DNA Data Storage via AHC and ECCs

Figure 5 indicates the influence of different single character lengths on storage density.

From Figure 5(a), the relationship between storage density and single character length is not linear but reaches a peak in a particular character length. This is because the number of required encoding symbols decreases with the increase of character length, but the types of symbols increase instead, reducing the probability of a character's occurrence. According to Figure 5(b), the AHC algorithm satisfies storage density for different storage files for 8-bit and 16-bit characters.

Under different insertion and deletion probabilities, the substitution error rate of DNA-ICC scheme is the bit error rate when Ps = 0, Ps = 0.1%, Ps = 0.2%, Ps = 0.3%, and Ps = 0.4%. The result is compared with that of other ECCs, such as **Bose ay-Chaudhuri Hocquenghem (BCH)** codes and Grid Matrix codes, as shown in Figure 6.

In Figure 6, the bit error rate of the DNA-ICC scheme decreases with error probability, meeting the changing trend of ECCs. The DNA-ICC scheme has the best performance compared with other error correction schemes. The error correction ability of Hamming Codes and RS Codes is poor. With the insertion error probability increase, the bit error rate is always high, and the downward trend is not apparent. Although the original Concatenated Code scheme can reduce the bit error

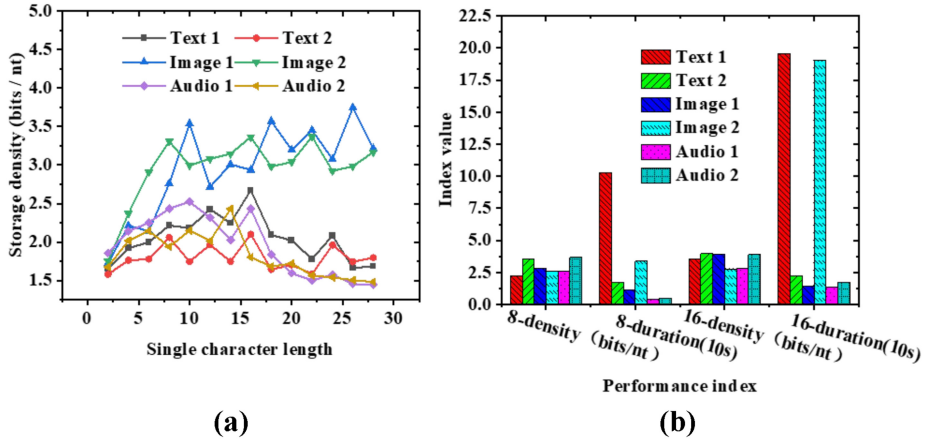

Fig. 5. Storage density under different character lengths: (a) storage density under different character lengths, (b) comparison of memory density and runtime for 8-bit and 16-bit characters.

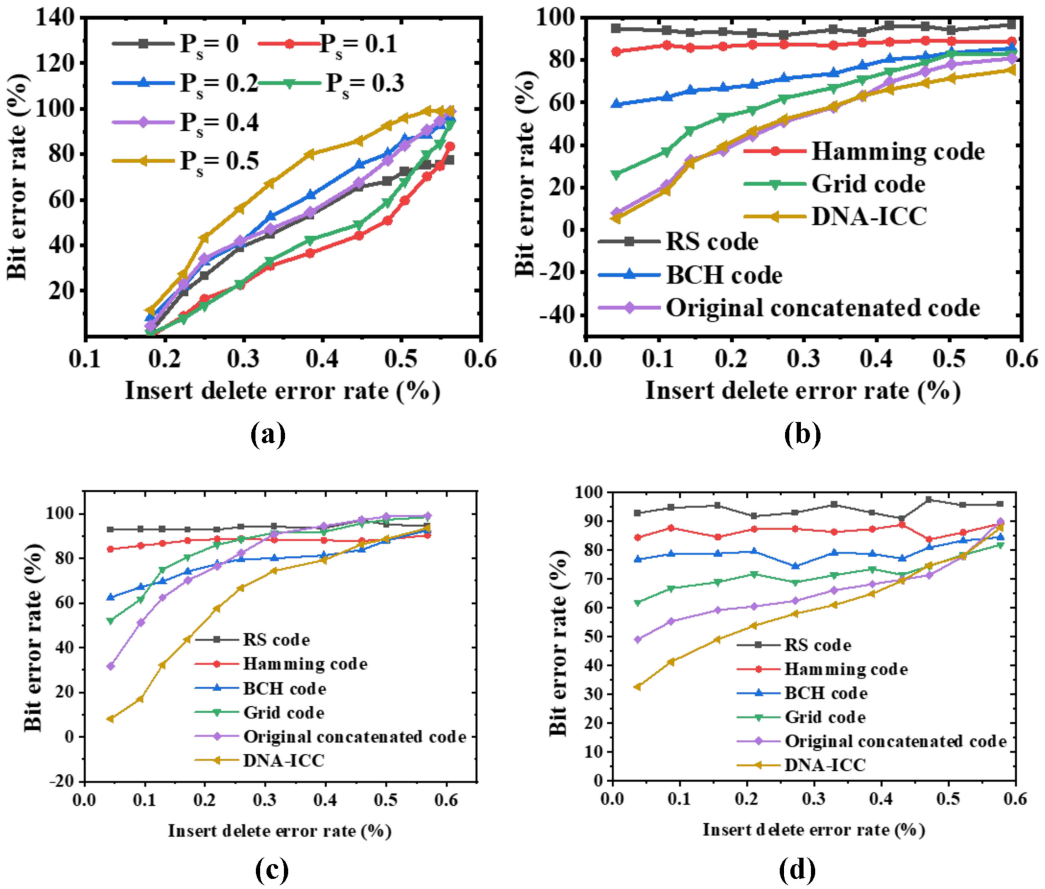

Fig. 6. Error rate comparison of different ECCs: (a) error rate comparison of different replacement error rates, (b) Ps = 0; (c) Ps = 0. 3; (d) Ps = 0. 4.

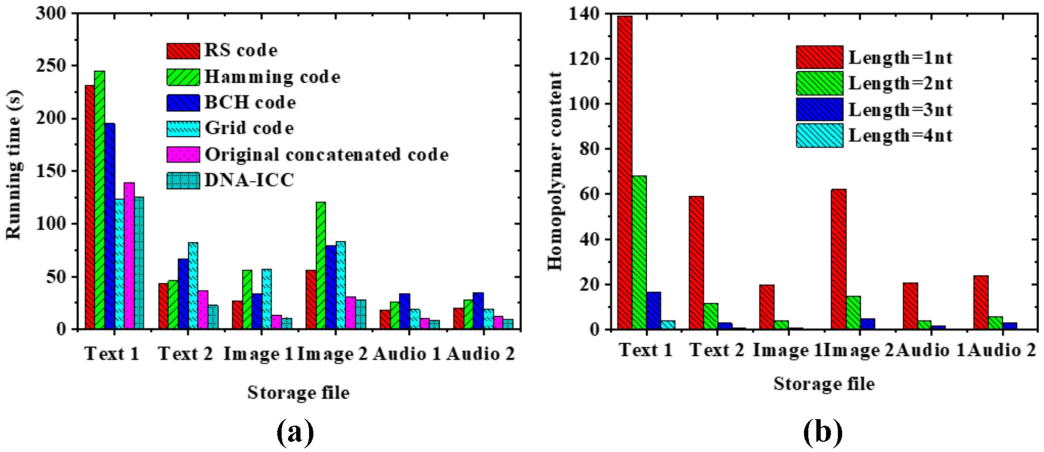

Fig. 7. Comparison of error correction time and Homopolymer of different files: (a) comparison of the running time of different error-correction schemes, (b) comparison of Homopolymer.

rate, the DNA-ICC scheme has a more robust error correction ability than the original Concatenated Code.

The error correction time and Homopolymer of the DNA-ICC scheme are shown in Figure 7.

Figure 7 demonstrates that the DNA-ICC scheme reduces encoding and decoding time and improves the efficiency of DNA information storage. Compared with the original Concatenated Code, it saves at least 1. 65s time. In addition, the DNA-ICC scheme has a good control effect on Homopolymer. When the size of Homopolymer exceeds 4nt, the occurrence probability of Homopolymer is as low as 0. 44%, close to 0.

5 CONCLUSION

With the development and broad application of 5G communication, IoT, cloud computing, big data, AI, and other new-generation information technologies, DT technology has rapidly developed both theory and application. It gradually extends to smart cities, parks, transportation, and other application fields. The DNA data storage model combining multi-source data storage and DNA computing in DT is gradually emerging. In this paper, six Degree Distribution function schemes are proposed for DNA information storage of Raptor Codes. This paper also improves Huffman Coding and the Cascaded ECCs, puts forward the quaternary adaptive Huffman encoding and decoding method, and optimizes the Concatenated Code. It is proved that the performance of Raptor Codes is greatly improved. However, there are some deficiencies in the research. Although the Raptor Code is a kind of information encoding and decoding with good performance, the encoding efficiency is still not high enough, less than 2bit/nt. Future research will consider neural networks and machine learning for optimization. Moreover, the storage density of hexadecimal AHC can be better. In addition, the future work will consider targeted coding for different file input types according to the specific content and structural characteristics of the input. It will also refer to the current communication or storage protocols to realize the mutual communication between DNA and computer data.

REFERENCES

[1] X. Li, B. Wang, H. Lv, Q. Yin, Q. Zhang, and X. Wei. 2020. Constraining DNA sequences with a triplet-bases unpaired. *IEEE Transactions on Nano-Bioscience* 19, 2 (2020), 299–307.

[2] K. A. S. Immink and K. Cai. 2020. Properties and constructions of constrained codes for DNA-based data storage. *IEEE Access* 8 (2020), 49523–49531.
[3] B. Cao, X. Zhang, J. Wu, B. Wang, Q. Zhang, and X. Wei. 2021. Minimum free energy coding for DNA storage. *IEEE Transactions on Nanobioscience* 20, 2 (2021), 212–222.
[4] B. Cao, S. Zhao, X. Li, and B. Wang. 2020. K-means multi-verse optimizer (KMVO) algorithm to construct DNA storage codes. *IEEE Access* 8 (2020), 29547–29556.
[5] J. H. Weber, J. A. De Groot, and C. J. Van Leeuwen. 2020. On eingle-error-detecting codes for DNA-based data storage. *IEEE Communications Letters* 25, 1 (2020), 41–44.
[6] K. J. Wang, Y. H. Lee, and S. Angelica. 2021. Digital twin design for real-time monitoring – a case study of die cutting machine. *International Journal of Production Research* 59, 21 (2021), 6471–6485.
[7] P. Major, G. Li, H. P. Hildre, and H. Zhang. 2021. The use of a data-driven digital twin of a smart city: A case study of Alesund, Norway. *IEEE Instrumentation & Measurement Magazine* 24, 7 (2021), 39–49.
[8] T. R. Wanasinghe, L. Wroblewski, B. K. Petersen, R. G. Gosine, L. A. James, O. De Silva... and P. J. Warrian. 2020. Digital twin for the oil and gas industry: Overview, research trends, opportunities, and challenges. *IEEE Access* 8 (2020), 104175–104197.
[9] F. Xue, W. Lu, Z. Chen, and C. J. Webster. 2020. From LiDAR point cloud towards digital twin city: Clustering city objects based on Gestalt principles. *ISPRS Journal of Photogrammetry and Remote Sensing* 167 (2020), 418–431.
[10] G. N. Schroeder, C. Steinmetz, R. N. Rodrigues, R. V. B. Henriques, A. Rettberg, and C. E. Pereira. 2020. A methodology for digital twin modeling and deployment in industry 4. 0. *Proceedings of the IEEE* 109, 4 (2020), 556–567.
[11] C. Zhang, G. Zhou, H. Li, and Y. Cao. 2020. Manufacturing blockchain of things for the configuration of a data-and knowledge-driven digital twin manufacturing cell. *IEEE Internet of Things Journal* 7, 12 (2020), 11884–11894.
[12] E. Yildiz, C. Møller, and A. Bilberg. 2020. Virtual factory: Digital twin based integrated factory simulations. *Procedia CIRP* 93 (2020), 216–221.
[13] W. Song, K. Cai, and K. A. S. Immink. 2020. Sequence-subset distance and coding for error control in DNA-based data storage. *IEEE Transactions on Information Theory* 66, 10 (2020), 6048–6065.
[14] K. Cai, Y. M. Chee, R. Gabrys, H. M. Kiah, and T. T. Nguyen. 2021. Correcting a single indel/edit for DNA-based data storage: Linear-time encoders and order-optimality. *IEEE Transactions on Information Theory* 67, 6 (2021), 3438–3451.
[15] X. Lu, J. Jeong, J. W. Kim, J. S. No, H. Park, A. No, and S. Kim. 2020. Error rate-based log-likelihood ratio processing for low-density parity-check codes in DNA storage. *IEEE Access* 8 (2020), 162892–162902.
[16] M. Campbell. 2020. DNA data storage: Automated DNA synthesis and sequencing are key to unlocking virtually unlimited data storage. *Computer* 53, 4 (2020), 63–67.
[17] J. Wang, L. Wang, J. Xu, and Y. Peng. 2021. Information needs mining of COVID-19 in Chinese online health communities. *Big Data Research* 24 (2021), 100193.
[18] P. Mishra, C. Bhaya, A. K. Pal, and A. K. Singh. 2020. Compressed DNA coding using minimum variance Huffman tree. *IEEE Communications Letters* 24, 8 (2020), 1602–1606.
[19] H. X. Nguyen, R. Trestian, D. To, and M. Tatipamula. 2021. Digital twin for 5G and beyond. *IEEE Communications Magazine* 59, 2 (2021), 10–15.
[20] Q. Qi, F. Tao, T. Hu, N. Anwer, A. Liu, Y. Wei... and A. Y. C. Nee. 2021. Enabling technologies and tools for digital twin. *Journal of Manufacturing Systems* 58 (2021), 3–21.
[21] C. Zhang, G. Zhou, H. Li, and Y. Cao. 2020. Manufacturing blockchain of things for the configuration of a data-and knowledge-driven digital twin manufacturing cell. *IEEE Internet of Things Journal* 7, 12 (2020), 11884–11894.
[22] J. Qian, B. Song, Z. Jin, B. Wang, and H. Chen. 2018. Linking empowering leadership to task performance, taking charge, and voice: The mediating role of feedback-seeking. *Frontiers in Psychology* 9 (2018), 2025.
[23] X. Li, B. Wang, H. Lv, Q. Yin, Q. Zhang, and X. Wei. 2020. Constraining DNA sequences with a triplet-bases unpaired. *IEEE Transactions on Nanobioscience* 19, 2 (2020), 299–307.
[24] Q. Zhou, X. Wang, and C. Zhou. 2021. DNA design based on improved ant colony optimization algorithm with bloch sphere. *IEEE Access* 9 (2021), 104513–104521.
[25] T. Xue and F. C. Lau. 2020. Construction of GC-balanced DNA with deletion/insertion/mutation error correction for DNA storage system. *IEEE Access* 8 (2020), 140972–140980.
[26] F. Xue, W. Lu, Z. Chen, and C. J. Webster. 2020. From LiDAR point cloud towards digital twin city: Clustering city objects based on Gestalt principles. *ISPRS Journal of Photogrammetry and Remote Sensing* 167 (2020), 418–431.
[27] S. Aheleroff, X. Xu, R. Y. Zhong, and Y. Lu. 2021. Digital twin as a service (DTaaS) in industry 4. 0: An architecture reference model. *Advanced Engineering Informatics* 47 (2021), 101225.
[28] L. Li, F. Gu, H. Li, J. Guo, and X. Gu. 2021. Digital twin bionics: A biological evolution-based digital twin approach for rapid product development. *IEEE Access* 9 (2021), 121507–121521.

[29] S. Namasudra, S. Sharma, G. C. Deka, and P. Lorenz. 2020. DNA computing and table based data accessing in the cloud environment. *Journal of Network and Computer Applications* 172 (2020), 102835.
[30] B. Adithya and G. Santhi. 2021. Deoxyribonucleic Acid (DNA) computing using Two-by-six complementary and color code cipher. *Bulletin of Computer Science and Electrical Engineering* 2, 1 (2021), 38–45.
[31] Y. Jiang, S. Yin, K. Li, H. Luo, and O. Kaynak. 2021. Industrial applications of digital twins. *Philosophical Transactions of the Royal Society A* 379, 2207 (2021), 20200360.
[32] S. Namasudra, R. Chakraborty, A. Majumder, and N. R. Moparthi. 2020. Securing multimedia by using DNA-based encryption in the cloud computing environment. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 3s (2020), 1–19.
[33] Y. Li, Y. Deng, X. Tang, W. Cai, X. Liu, and G. Wang. 2018. Cost-efficient server provisioning for cloud gaming. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 3s (2018), 1–22.
[34] S. Kuruppu, B. Beresford-Smith, T. Conway, and J. Zobel. 2011. Iterative dictionary construction for compression of large DNA data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9, 1 (2011), 137–149.
[35] S. Hassantabar, N. Stefano, V. Ghanakota, A. Ferrari, G. N. Nicola, R. Bruno... and N. K. Jha. 2021. CovidDeep: SARS-CoV-2/Covid-19 test based on wearable medical sensors and efficient neural networks. *IEEE Transactions on Consumer Electronics* 67, 4 (2021), 244–256.
[36] Z. Beini, C. Xuee, L. Bo, and W. Weijia. 2021. A new few-shot learning method of digital PCR image detection. *IEEE Access* 9 (2021), 74446–74453.
[37] K. Lischer, A. B. R. Digdaya Putra, M. Sahlan, A. C. Khayrani, M. J. Ginting, A. Wijanarko... and D. K. Pratami. 2021. Heat transfer simulation of various material for polymerase chain reaction thermal cycler. *Journal of Mechanical Engineering (JMechE)* 8, 2 (2021), 27–37.
[38] H. Zheng, J. Wang, J. Zhang, and R. Li. 2021. IRTS: An intelligent and reliable transmission scheme for screen updates delivery in DaaS. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 3 (2021), 1–24.

Received 30 January 2022; revised 29 June 2022; accepted 1 September 2022