Efficient DNA-Based Image Coding and Storage

Cihan Ruan*, Rongduo Han†, Yixiao Li‡, Shan Gao§, Haoyu Wu¶, and Nam Ling*
cruan@scu.edu a12910@qq.com doveliyixiao@163.com gao_shan@mail.nankai.edu.cn wuhaoyujerry@gmail.com nling@scu.edu
*Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA, USA
†School of Mathematical Sciences, Nankai University, Tianjin, China
‡Industrial and Commercial Bank of China, Beijing, China
§College of Life Sciences, Nankai University, Tianjin, China
¶Roku, Inc., San Jose, CA, USA

2023 IEEE International Symposium on Circuits and Systems (ISCAS) | 978-1-6654-5109-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCAS46773.2023.10182162

***Abstract*—As global data volume explodes, traditional storage systems face multiple challenges, including the lack of resources and low cost-efficient. Deoxyribonucleic acid (DNA) has attracted researchers' attention as a novel storage medium to address these issues with its high storage density, low maintenance cost, and extremely long shelf life. Specifically, DNA is also environmentally friendly compared to traditional disk storage because the disk is non-degradable. In this paper, we propose a strategy for image coding and storage based on compressing intra-predicted images from Versatile Video Coding (VVC) using synthetic genomics theories that enable the high throughput storage of images on DNA. We first define the length and format of the DNA oligo for the implementation of a storage system. Then we improve the LT codes to become a feasible DNA coding scheme. Last but not least, we design a voting mechanism to achieve the error correction function for robustness. The experimental results show high compression efficiency while achieving several strict biological constraints, such as GC-content balance and homopolymer control.**

***Index Terms*—Image coding, visual communications, Versatile Video Codec, GC-content balance, homopolymer control, synthetic genomics, high throughput DNA storage**

I. Introduction

According to the International Data Corporation (IDC)'s forecast, the total amount of digital data worldwide is expected to reach 175 Zettabytes (ZB) in 2025 [1] because of the prevalence of social media in multimedia format. There are laws in multiple countries that require this data to be archived for a long time even though it may not be accessed often. With an explosive growth of global data volume, traditional storage media, like magnetic, optical, and flash archival recorders [2], can no longer meet the current high demand for data storage. Firstly, the materials of traditional storage media are not durable. Their lifetimes are estimated from 3 to 50 years. Secondly, data stored can be easily corrupted, hence requiring a lot of labor to maintain data security. Lastly, traditional storage media have low storage density. As a result, a novel storage medium needs to be adopted to address the increasing global volume.

Deoxyribonucleic acid (DNA) sequences can be considered innovative natural storage materials. In particular, DNA is encoded by the arrangement of four kinds of nitrogenous bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Based on recent technology in synthetic genomics, DNA can be artificially created in a specific order of bases to store information. Compared to hard storage, which can only be stored for decades and require a lot of energy and human resources, the processed DNA strands can be stored in small test tubes for thousands of years in a room-temperature environment [3]. Moreover, DNA has a very high storage density, which is about 1 exabyte/mm^3 (109 GB/mm^3) [4]. Because of these characteristics, DNA is an ideal and environmentally friendly solution for large-scale data archival storage.

The research on DNA storage can be traced back to the 1970s [5]. Still, due to the limitation of technologies in DNA synthesis and sequencing, DNA storage is more like an ambitious blueprint rather than a usable storage system for a long time. With the development of synthetic genomics in the last decade, DNA storage has come back to researchers' minds. Theoretically, the main encoding process of DNA storage usually contains the following steps:

1) Designing a coding scheme corresponding to four nitrogenous bases.
2) Converting the desired input data to be stored from binary to quaternary DNA strands format.
3) Sending the result over to the biology lab for a series of wet lab experiments, such as DNA synthesis, polymerase chain reaction (PCR) amplification, and DNA sequencing.

In reality, there are more difficulties than that. For example, an efficient data compression algorithm is required because DNA synthesis cost is expensive. The more information we can store per base, the lower cost we need to spend. Plus, the arrangement of bases can be arbitrary. According to synthetic genomics, there are several biological constraints that we need to observe when designing DNA synthesis strands:

- Homopolymers Control: Avoiding consecutive occurrences more than three times.
- GC-content Balance: Controlling the GC-content, the percentage of nitrogenous bases in each DNA sequence, including guanine (G) or cytosine (C), should be in the range of 40%-60% [6].
- Repetitions Forbidden: Avoiding repetitive patterns when designing coding mechanisms [7].

In addition, DNA storage has its own disadvantages: inevitable

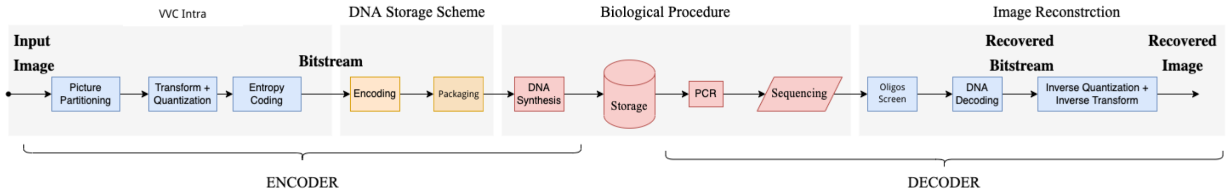

Fig. 1. The proposed DNA coding workflow

random errors caused by PCR and sequencing, slow I/O speeds, and the inability to directly support random I/O access. All of these need to be considered when designing the storage architecture.

Multiple researchers have confirmed that we could successfully store small data in artificial DNA strands and recover it [8]–[10]. They provided different kinds of algorithms for DNA storage with various redundancies and compression densities. Some works also contributed to the robustness and random I/O function as disk storage by designing the DNA strand structure. Based on these achievements, researchers started to save larger images and even videos that adopted different compression schemes. In 2020, Pan et al. published an algorithm that converts RGB images to one-dimensional bitstreams by color channel and then to DNA base strings using Huffman coding [11]. Li et al. enhanced DNA storage robustness by refining the DNA oligo design structure in 2021 [12]. Later, Dimopoulou et al. proposed a JPEG-DNA storage mechanism based on the original JPEG standard but improved the 0-1 binary Huffman coding embedded in JPEG encoding to A-G-C-T encoding [13]. However, those studies were based on earlier image compression algorithms with limited coding efficiency.

Versatile video codec (VVC) or H.266/MPEG-I (ISO/IEC JTC 1) is the latest coding standard in the video codecs domain. Its intra-coding part can be treated as one of the latest image codecs [14]. VVC (intra) adopts a hybrid coding framework consisting of prediction, transformation, quantization, and entropy coding. Compared to the previous generation, High Efficiency Video Coding (HEVC), VVC can improve compression efficiency by 30%-50% [15], specifically suitable for HD and 4K image compression.

In this paper, we propose a new high-density encoder to store high-quality images with the error correction capability of DNA oligos for VVC intra-coded images. We present algorithmic content and details in section II and the implementation of the proposed algorithm in section III. Finally, we conclude our work and discuss the future in section IV.

II. Proposed Method

With the great compression ratio of VVC (intra), we can compress images into bitstreams. Then we convert the bitstreams to several single short DNA strands called oligos. Those oligos that meet biological constraints, can be used for synthesis and stored in an appropriate environment. When decoding, oligos are decoded back to binary bitstreams. We can recover the bitstreams to images with a VVC (intra) decoder and proper error correction mechanism.

The workflow of the proposed DNA storage system based on VVC (intra) is shown in Fig. 1. Related technical details are described below.

A. Oligo Length

The main issue limiting the development of DNA storage is the development of synthesis and sequencing technologies. Using a long DNA strand to implement a DNA storage system is inapplicable. We set the maximum length of oligos as 300 nt to achieve a high throughput DNA storage system according to the current strategy of the most efficient platform of synthesis and sequencing provided by *Agilent SurePrint* and *Illumina* [16].

B. LT Codes

Based on Shannon's information capacity theory, the maximum information capacity in DNA storage is two bits for each nitrogenous base. During the storage process, the data needs to be split into small segments for further operation. Since the fountain code can recover the whole data by taking only some of the small segments. Plus, it has high density. In this case, we choose the fountain code as the baseline.

Luby transform codes (LT codes) are the most practical among fountain code algorithms. Inspired by DNA Fountain [17], we adopt LT codes [18] as the prototype but made several improvements.

C. Oligo Format

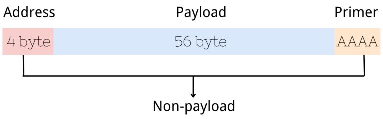

Fig. 2. The design of DNA oligo format

Considering the coding potential of LT codes, we define approximately 60 bytes of the data segment to be mapped and stored in a 300 nt DNA oligo. However, each DNA oligo designed for synthesis needs to be built-in associate primers taken with several bases. Besides primers, 4 bytes are used to store the seed address of the LT transform, and the rest 56 bytes are stored as the payload of this DNA oligo. In some related work, a non-payload section for error correction is designed. We have also designed the voting mechanism for error correction to further increase the compression ratio. The format of our desired DNA oligo is shown in Fig.2.

D. The Coding Details

Our improved LT encoder consists of three main parts: packaging, Luby transform, and adaptive DNA mapping based on the context (Fig. 3). We also design a voting verification algorithm for error correction on the decoding side.

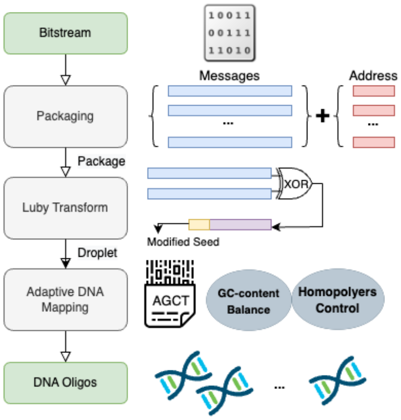

Fig. 3. The encoder workflow

1) Packaging: Because the bitstream generated from VVC (intra) is well compressed, we can directly partition the bitstream of each image into non-overlapping data segments and package them. Here, we set the size of each slice as 56 bytes. Plus, we made an additional header oligo that contains metadata. We assume the total number of generated packages is *M*.

2) Luby Transform and Modified Seed Algorithm: Luby transform selects random packages from the previous step and applies bitwise addition. Then it attaches a random seed and generates various desired numbers of short messages named droplets. We mark the number of fountain code droplets as *N*. The fixed-length seed in the original algorithm is created by both the numbers of total generated packages and the current package's attribute. Such a scheme requires the storing of two values and takes up more space, so we improve it by using only one index to generate all the messages. We use *M* as a seed and generate a sampling distribution of length *2M*. Because that *2M* might be smaller than *N*, so we calculate $i \mod 2M$ as the indicator to get the number of original packages in the corresponding droplet. *M* is stored as a unique parameter in the header DNA oligo.

3) Adaptive DNA Mapping: In this step, DNA Fountain converts the binary droplet to the DNA oligo by using a simple binary-quaternary coding Rule and then discards those oligos that do not meet biological constraints.

TABLE I
A SAMPLE CODEBOOK FOR THE ADAPTIVE DNA MAPPING

	AT Code book				CG Code book			
A	TA	TC	TT	TG	CA	GC	CT	CG
C	TA	TC	AT	TG	GA	GC	GT	GG
T	AA	AC	AT	AG	GA	GC	GT	CG
G	AT	AC	TA	TG	CA	CC	CT	CG

To save computing resources, we design an adaptive mapping algorithm based on the preceding DNA to improve the LT codes. Instead of generating many sequences and screening them, we directly generate sequences that meet the requirements. We record the last two bases coded and calculate the GC-content of the previous text. We also generate *dynamic codebooks*, which create a mapping scheme for the next bit based on the context and global GC-content. Specifically, we can set a threshold value for *GC-content control* as a limitation. If the GC-content is higher than the threshold, the amount of AT in the dynamic codebooks can be increased. The increment is determined by the current global CG-content of the encoded oligo, and we increase the AT-content by 5 quantiles (e.g. 0%, 25%, 50%, 75%, 100%). When the first two bases are equal, a specific codebook needs to be created to ensure *homopolymers control* (with a certain probability p), to make sure there are no three consecutive bases. The next bit n can be determined by the length l of contiguous bases in the oligo and the error of p (Eq. 1). Moreover, to further improve the compression rate, the codebook needs to ensure that the desired length of mapping bases is as small as possible.

$$f\left(n\right) \rightarrow \{A, C, T, G, AA, AC...\}$$
$$Length = \sum_{n\epsilon 0,1,2,3} p(n) l(f(n)) \tag{1}$$

Since there are consecutive repetitions in the data, it is necessary to increase the randomness of the codebooks to ensure that a more diverse and highly compressed DNA oligo can be created (Eq. 2). R_1 and R_2 are random primes to generate pseudo-random numbers. A sample of codebooks is shown in TABLE I.

$$Offset = \left[\frac{Length(Encoded))}{R_1}\right] \mod R_2 \tag{2}$$
$$f(n\epsilon\left\{0,1,2,3\right\}) = \left\{A, C, T, G\right\}\left[n_{Location} + Offset\right]$$

By designing the above mechanism, our proposed algorithm eliminates the screening steps and reduces time complexity compared to relative works.

4) Robustness: DNA oligos are amplified and sequenced by PCR and can be decoded back into bitstream format. Usually, we need to justify whether the DNA oligo and the recovered information are incorrect or corrupted. Some researchers try to insert error correction code, but it causes redundancy. In DNA storage, PCR amplification creates thousands of duplicates of one oligo. Since we already have duplicate oligos, we choose to use a voting method for calibration. In most conditions, more than 98% of the DNA bases are correct for a 300 nt oligo. According to Hoeffding's inequality, with the increasing number of created oligos N, the correct probability p converges exponentially to 1 as shown in Eq.3. Therefore, the voting scheme can provide reliable outcomes.

TABLE II
RESULTS OBTAINED BY THE PROPOSED ENCODER ON THE TEST DATASET
(DPI: DOTS PER INCH, QP: QUANTIZATION PARAMETER, PACKAGES: NUMBER OF PARTITION MESSAGES, NT: DNA BASE UNIT)

Image	Resolution/dpi	QPs	PSNR/dB	Packages	Droplets	Ave. Oligo Len/nt	A-content	C-content	G-content	T-content
$Girl$	768x512	28	40.4481	361	722	292.2632	0.26	0.25	0.26	0.23
Hat	768x512	28	42.1326	275	550	292.1836	0.26	0.25	0.26	0.23
$Building$	768x512	23	42.5871	1640	3280	292.2323	0.26	0.25	0.26	0.23
$Parrot$	768x512	28	42.3262	245	490	292.2796	0.26	0.24	0.26	0.24
$Lighthouse$	512x768	23	43.0601	938	1876	292.3001	0.26	0.25	0.26	0.23
$Restaurant$	3054x2336	23	40.7064	16581	33162	291.6051	0.26	0.24	0.26	0.24
$Cafe$	3872x2592	23	39.7179	22779	45558	291.5826	0.26	0.24	0.27	0.23
Dog	4256x2832	28	44.4834	1113	2226	292.1765	0.26	0.25	0.26	0.23

TABLE III
COMPARISON OF ADDITIONAL BINARY-QUANDARY ENTROPY CODING WITH RELATED WORKS

Refs	Oligo Len/nt	Bits/nt	Coding Scheme	Robustness
Goldman [8]	115	0.33	Rotating Encoding	Repetition
Church [7]	117	0.83	1bit to 1 base	No
Organick [20]	150-200	1.10	Reed-Solomon Coding	Error Correction Code
Our Work	300	1.8	Semi-LT codes+Dynamic Mapping	Voting

$$P\left[H(N) \leq \left[\frac{N}{2}\right]\right] = \sum_{i=0}^{\frac{N}{2}} \mathrm{C}_n^i p^i (1-p)^{N-i} \leq e^{-2(\frac{2p-1}{2})^2 N} \quad (3)$$

In addition, since we define the size of packages as a fixed value and the wrong sequences lead to packages with incorrect lengths when decoding, which can be discarded directly before voting. To further reduce the operational complexity, we apply the MD5 message-digest algorithm to calculate the MD5 hash value for each sequence [19]. Sequences with the same hash value are divided into groups for voting. This scheme provides the guarantee of robustness.

III. EXPERIMENTAL RESULT

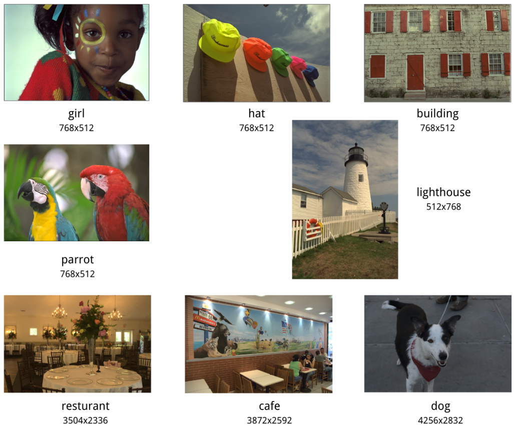

Fig. 4. Test Image Set

We selected eight images of different resolutions to compose a test image set (Fig. 4). The results of our coding algorithm are shown in TABLE II. Our improved LT codes as an additional binary-quandary entropy coding achieve an encoding efficiency of 1.8 bits/nt, which is close to Shannon's theoretical maximum value and higher than the other works (TABLE III). When we combined VVC (intra) with the proposed coding scheme, our method also achieved an impressive overall logical storage density regardless of the image resolutions (Fig. 5) while maintaining a low redundancy and a high peak-signal-to-noise-ratio (PSNR). Compared with other DNA image storage algorithms, e.g. 4.9708 bit/nt in JPEG-DNA [13], our work had a superior result. This achievement was partly due to the contribution of VVC, for example, the image "*dog*" received an excellent compression result because of its simple background, but our algorithm further improves the results. In addition, our algorithm reduces complexity compared to other fountain codes. This improvement offers a foundation to handle large-scale data and achieve high throughput storage.

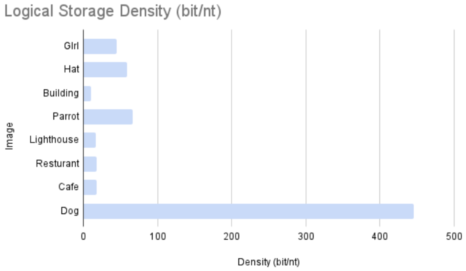

Fig. 5. Logical storage density of test images

IV. CONCLUSION

This paper proposes a robust high-definition image DNA storage coding scheme and applies it to VVC intra coded images. We obtain a highly compressed bitstream from VVC (intra) and improve the LT codes to convert the binary bitstream to quaternary nitrogenous bases of DNA with biological constraints. Our proposed method guarantees a high information capacity of the code, saves computing power, and reduces data redundancy. From experiments, compared with other related work, our overall compression density is much higher. Our work provides an excellent theoretical basis for future high-throughput, large-scale information storage in DNA.

References

[1] J. G. David Reinsel and J. Rydnin, "The digitization of the world from edge to core," *IDC White Paper*, 2018.

[2] K. Goda and M. Kitsuregawa, "The history of storage systems," *Proceedings of The IEEE - PIEEE*, vol. 100, pp. 1433–1440, 05 2012.

[3] N. V. Ivanova and M. L. Kuzmina, "Protocols for dry DNA storage and shipment at room temperature," *Molecular ecology resources*, vol. 13, no. 5, pp. 890–898, 2013.

[4] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-based archival storage system," in *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, 2016, pp. 637–649.

[5] J. Davis, "Microvenus," *Art Journal*, vol. 55, no. 1, pp. 70–74, 1996.

[6] S. Tabatabaei Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Scientific reports*, vol. 5, no. 1, pp. 1–10, 2015.

[7] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[8] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *nature*, vol. 494, no. 7435, pp. 77–80, 2013.

[9] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, 2015.

[10] M. Blawat, K. Gaedke, I. Huetter, X.-M. Chen, B. Turczyk, S. Inverso, B. W. Pruitt, and G. M. Church, "Forward error correction for DNA data storage," *Procedia Computer Science*, vol. 80, pp. 1011–1022, 2016.

[11] C. Pan, S. H. T. Yazdi, S. K. Tabatabaei, A. G. Hernandez, C. Schroeder, and O. Milenkovic, "Image processing in DNA," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8831–8835.

[12] B. Li, L. Ou, and D. Du, "IMG-DNA: approximate DNA storage for images," in *Proceedings of the 14th ACM International Conference on Systems and Storage*, 2021, pp. 1–9.

[13] M. Dimopoulou, E. G. San Antonio, and M. Antonini, "A jpeg-based image coding solution for data storage on DNA," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 786–790.

[14] S. N. Kumar, M. V. Bharadwaj, and S. Subbarayappa, "Performance comparison of jpeg, jpeg xt, jpeg ls, jpeg 2000, jpeg xr, hevc, evc and vvc for images," in *2021 6th International Conference for Convergence in Technology (I2CT)*. IEEE, 2021, pp. 1–8.

[15] O. Watanabe, R. Suzuki, and H. Kiya, "A structure of jpeg xt encoder considering effect of quantization error," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2016, pp. 810–813.

[16] Q. Qian, L. Bo, Y. Lin, W. Bing-bing, W. Hui-jun, D. Xin-Ran, L. Yu-lan, and Z. Wen-Hao, "Application of copy number variation screening analysis process based on high-throughput sequencing technology," *Chinese Journal of Evidence-Based Pediatrics*, vol. 13, no. 4, p. 275, 2018.

[17] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *science*, vol. 355, no. 6328, pp. 950–954, 2017.

[18] M. Luby, "Lt codes," in *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.* IEEE Computer Society, 2002, pp. 271–271.

[19] R. Rivest, "The md5 message-digest algorithm," Tech. Rep., 1992.

[20] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen *et al.*, "Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, pp. 242–248, 2018.

[21] S. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Scientific reports*, vol. 7, no. 1, pp. 1–6, 2017.

[22] M. Antonini, L. Cruz, E. da Silva, M. Dimopoulou, T. Ebrahimi, S. Foessel, E. G. San Antonio, G. Menegaz, F. Pereira, X. Pic *et al.*, "DNA-based media storage: State-of-the-art, challenges, use cases and requirements version 7.0," 2022.

[23] B. Cao, X. Zhang, S. Cui, and Q. Zhang, "Adaptive coding for DNA storage with high storage density and low coverage," *NPJ systems biology and applications*, vol. 8, no. 1, pp. 1–12, 2022.