Towards Practical and Robust DNA-based Data Archiving by Codec System Named 'Yin-Yang'

Zhi Ping[1,3,4,†], Shihong Chen[1,3,5,†], Guangyu Zhou[6,7,†], Xiaoluo Huang[1,†], Sha Joe Zhu[8], Chen Chai[1,4], Haoling Zhang[1,3,4], Henry H. Lee[6], Tsan-Yu Chiu[2,4], Tai Chen[2,4,5], Huanming Yang[1,4], Xun Xu[2,5,*], George M. Church[4,6,9,*], Yue Shen[1,3,4,5,*]

†These authors contributed equally to this work.

*To whom correspondence should be addressed (shenyue@genomics.cn, gchurch@genetics.med.harvard.edu, and xuxun@genomics.cn)

[1]Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Guangdong, China

[2]Guangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Guangdong, China

[3]Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, People's Republic of China

[4]George Church Institute of Regenesis, BGI-Shenzhen, Shenzhen, China

[5]China National GeneBank, BGI-Shenzhen, Shenzhen, China

[6]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.

[7]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, United States.

[8]Big data Institute, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, Oxford, United Kingdom.

[9]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts, USA.

Abstract

Motivation: DNA has been reported as a promising medium of data storage for its remarkable durability and space-efficient storage capacity. Here, we propose a robust DNA-based data storage method based on a new codec algorithm, namely 'Yin-Yang'. **Results:** Using this strategy, we successfully stored different file formats in a single synthetic DNA oligonucleotide pool. Compared to most well-established DNA-based

data storage coding schemes presented to date, this codec system can achieve a variety of user goals (e.g. reduce homopolymer length to 3 or 4 at most, maintain balanced GC content between 40% and 60% and simple secondary structure with the Gibbs free energy above -30 kcal/mol). It also shows enhanced robustness in transcoding of different data structure and practical feasibility. We tested this codec with an end-to-end experiment including encoding, DNA synthesis, sequencing and decoding. Through successful retrieval of 3 files totaling 2.02 Megabits after sequencing and decoding, our strategy exhibits great qualities of achieving high storing capacity per nucleotide (427.1 PB/gram) and high fidelity of data recovery.

Introduction

DNA is the most ancient information carrier designed by nature. At present, it is considered to have great potential to be a novel storage medium, as the orthodox storage media can no longer meet the exponentially increasing data archiving demand. DNA molecule, compared to orthodox information carriers, exhibits multiple advantages; including extremely high storage density (455 EB or 455×10^{18} bytes per gram theoretically[1]), extraordinary durability (an estimated half-life of DNA in bone was 521 years[2]), and the capability of cost-efficient information amplification.

Multiple strategies were put forward for digital information storage using organic molecules, including oligopeptides[3] and metabolomes[4]. Recently, 'DNA memory' using DNA origami technology was also proposed to be a novel strategy to store digital information. However, owed to the low cost and high throughput of current DNA sequencing instead of mass spectroscopy or super-revolution microscopy technologies, data archiving using synthetic DNA is still the most practical and cost-efficient strategy in this field. Binary information from each file is transcoded directly into DNA bases, synthesized, the polymer is then stored in the form of oligonucleotide or fragment. Specific patterns in DNA sequences are difficult for accurate synthesis and sequencing, e.g. abnormal GC content or long single-nucleotide repeats (homopolymers)[5-7]. Therefore, a transcoding algorithm should strive to maximize the coding potential (bits/nt) while preserving synthesis and sequencing capabilities. In the past few years, several different coding schemes were developed to achieve maximize coding potential. Some well-established algorithms, including Church's 'simple code'[1], Goldman's code[8], Grass' code[9] and DNA fountain[10], have sought to preserve synthesis and sequencing

capabilities and data fidelity by reducing homopolymers, balancing GC content, while allowing for the introduction of error-correction codes and other redundancies[11]. In addition, random-access is a necessary technique in DNA-based data storage, it was reported that several different molecular strategies can be used to achieve this function[7, 12, 13].

In order to achieve robust DNA-based data storage with high data density as well as reliability, we propose a novel binary-DNA codec algorithm and demonstrate its application in DNA-based data storage. It is named after a traditional Chinese concept, 'Yin-Yang'. 'Yin' and 'Yang' represent two different coding rules, which are incorporated subsequently to transcode binary information to sequences of 'A/T/C/G'. This codec algorithm can effectively reduce long homopolymers and maintain balanced GC content with reduced secondary structures for synthesized DNA sequences. Importantly it is capable of achieving the theoretical maximum DNA coding potential of ~2 bits/nt.

The general principle of 'Yin-Yang' codec algorithm (referred as 'YYC' hereinafter) is to overlap two independent encoding rules, 'Yin' and 'Yang', into one DNA sequence (referred as 'incorporation' hereinafter), thereby 'compressing' two bits into one nucleotide. Both rules can derive to 256 and 6 different specific rules respectively (Supplementary information). The incorporation of these two rules can derive a pool of 1536 different ways to encode the binary data and thus be possible to broaden the applications (Fig. 1A, Supplementary Fig. S1).

As an example (referred as Rule No. 888 in Figure. 1B), in 'Yang' rule, {'A', 'T'} represents binary digit '0' and {'G', 'C'} represents binary digit '1'. In 'Yin' rule, the local nucleotide (current nucleotide to be encoded) is represented by the incorporation of the previous nucleotide (or 'supporting nucleotide') and the corresponding binary digit (Fig. 1B). In practice, the two rules are applied respectively for two independent binary segment and transcoded into one exclusive DNA sequence, and vice versa. Taking the first nucleotide in Fig. 1C as a simple example, the corresponding binary digits from segment 'a' and 'b' are '1' and '0' respectively. Therefore, according to 'Yang' rule, the local nucleotide will have two options, {'C', 'G'}. On the other hand, according to 'Yin' rule, the virtual supporting nucleotide is set as 'A' by default, so previous nucleotide 'A' and corresponding binary digit '0' will give the local nucleotide

two options, {‘A’, ‘G’} (Fig. 1B). Therefore, the result encoded local nucleotide will be the intersection from ‘Yang’ rule, {‘C’, ‘G’} and ‘Yin’ rule, {‘A’, ‘G’}, which provides the only option {‘G’}, and so on. Moreover, it is also crucial to pay attention to the corresponding rules of each segment, because the switch of binary segments will also change the transcoded result, which means, {a: ‘Yang’, b: ‘Yin’} and { b: ‘Yang’, a: ‘Yin’} will result in completely different result in this case. The simple example of transcoding demonstration process is also illustrated in supplementary information (Fig. 1C, Supplementary video).

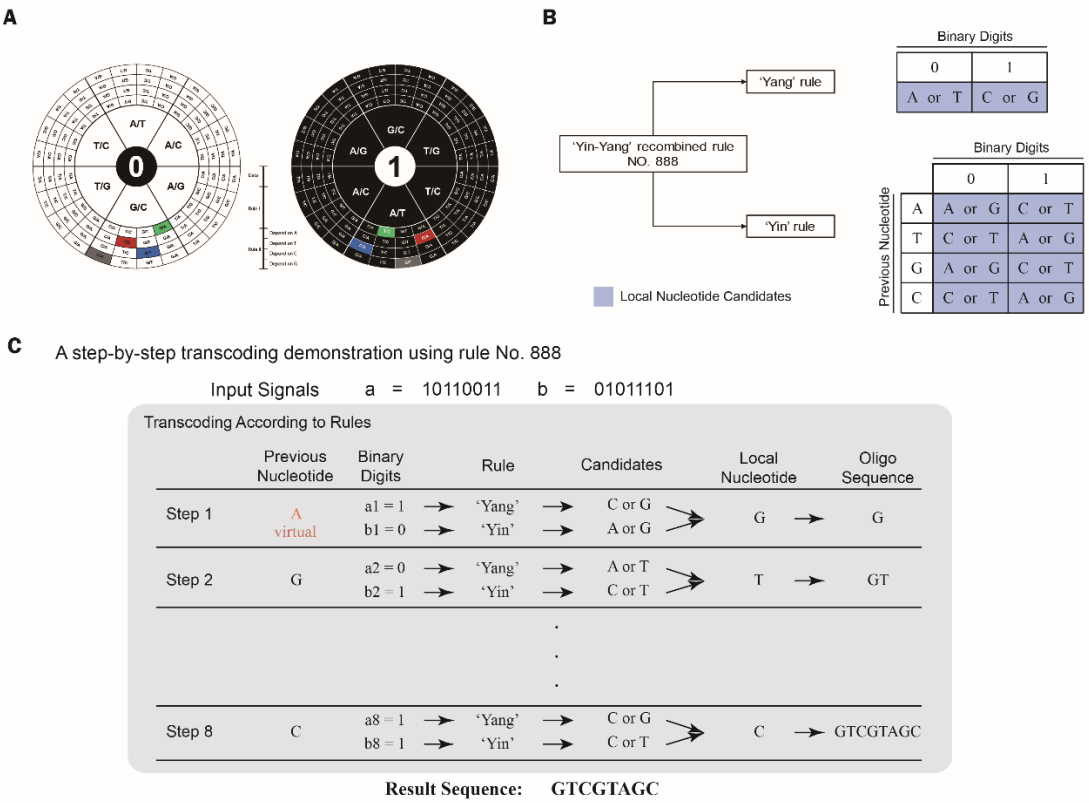


Figure 1 (A) ‘Yin-Yang’ incorporation rules can provide a pool of 1536 (6×256) different incorporated rules. See Supplementary Information for details. (B) An illustration of ‘Yin-Yang’ incorporated rule No. 888. (C) A step-by step demonstration of YYC transcoding process, rule No. 888 is used in this example. ‘a1’, ‘b1’ represents the first-position binary digit in segment ‘a’ and ‘b’, and so on. Virtual base A means this base is used only for determination of the first output base and will not be appear in the result sequence.

Results

High coding potential

Coding potential is always one of the key features of DNA-based data storage coding scheme evaluation. For current DNA codec algorithms based on natural nucleic acids (Adenine, Cytosine, Guanine and Thymine), the theoretical limit of coding potential is 2 bits/nt according to Shannon's information theory[14]. The novel strategy of 'Yin-Yang' codec algorithm (Methods) provides an extremely high coding potential. We show that 'Yin-Yang' codec exhibits the highest information density compared to current coding algorithms under identical constraints, including oligo length, index, flanking sequence, error-correction code and copy-number of oligos for DNA storage (Table. 1). In practical, logical redundancy using XOR operation[12, 15] was introduced in case of oligo loss, resulting a real information coding density of 1.27 bits/nt, while the proportion of index, error correction and primer bases (non-data region) in one oligo was $\frac{14+8+20\times2}{182} = 34\%$ (Fig. 3B) in this demonstration. With increasing length of synthesized DNA while maintaining the length of the non-data region, the coding potential reach the theoretical limit of 2 bits/nt (Supplementary Fig.S3). Beside theoretical information density, we also calculated the physical density using different encoding algorithms. The physical density (Bytes/gram) can be calculated through the following formula:

$$\rho = n \times (M\ molecules \times \frac{200nt}{molecule} \times \frac{320Dalton}{nt} \times \frac{1.67 \times 10^{-24}g}{Dalton})^{-1},$$

where n is the number of bytes each oligo carries, and M is the average copy number of each oligo. In previous working scheme[10], the minimum average copy number for data retrieval was around 1.3×10^3. In this work, we demonstrated that the minimum average copy number for data retrieval is 7×10^2 (Supplementary information). Therefore, we normalized all calculations of physical density to an average copy number of 1300 and 700 for comparison (Table. 1).

Table 1 Comparison of information efficiencies of current coding algorithms used in DNA-based data storage. The information density is calculated with identical parameters: length of oligo length (200nt), index (16nt), flanking sequence (24nt×2), error-correction code (8nt) and copy-number of oligos in DNA synthesis (700 copies per oligo type).

Coding Algorithm	***Information density***	***Physical density PB/g (Copy* No. =**	***Physical density PB/g (Copy* No. =**	***Theoretical Coding Potential***

		700)	*1300)*	
Simple code	0.64 bits/nt	213.9	115.2	1 bit/nt
Goldman *et. al*	1.008 bits/nt	336.8	181.4	1.58 bits/nt
Grass *et. al*	1.138 bits/nt	380.3	204.8	1.78 bits/nt
Blawat *et. al*	1.024 bits/nt	342.2	184.3	1.6 bits/nt
DNA Fountain	1.196 bits/nt	400.0	215.2	1.98 bits/nt*
'Yin-Yang' code	1.28 bits/nt	427.1	230.3	2 bits/nt

*** based on length of seed.**

Comparisons between YYC and current well-established coding schemes

To prove the feasibility of 'Yin-Yang' codec algorithm and quantify its featured parameters compared to other well-established DNA-based data storage coding schemes[1, 8-10], a collection of different formats of files including text, image, audio and video files, with the total size of ~1 GB, were transcoded using different algorithms *in silico*. Statistics of GC content, free energy of secondary structure, and data retrieval efficiency based on the number of read DNA oligo, are obtained and analyzed. For previous coding schemes except for DNA Fountain, most of the DNA oligos possess relatively normal GC content (40% - 60%) (Fig. 2A). However, sequences with high or low GC ratio still exist, which might cause difficulties for synthesis and sequencing. In contrast, YYC and DNA fountain can maintain the GC content in certain range (40% - 60% in this work, Fig. 2B). Because in both coding schemes, the incorporation of binary segments is flexible, which provides more possibilities to obtain DNA sequences with desired GC content. Furthermore, a procedure called 'validity-screening' can be

optionally chosen and used to help filter DNA sequences which do not meet other design criteria (Supplementary Fig. S2).

A new feature of YYC that is not found in previous codec is to predict oligo secondary by calculating the sequence free energy (Fig. 2C). Based the empirical data of DNA supplier from BGI, oligos with predicted free energy less than -30 kcal/mol tend to form more stable secondary structure and may cause troubles for molecular manipulation such as PCR amplification.

A linear retrieval pattern in which the amount of recovered information is proportional to the number of sequences read[11] would be desirable. YYC is capable of increasing coding potential while maintaining linear retrieval. In contrast, DNA Fountain, which employs the algorithm originally designed for communication channel, uses a different data retrieval strategy based on its grid-like topology of data segments, and will only recover all the data if enough segments are read[10]. Therefore, if an insufficient number of oligos are sequenced, there will be a shortage of data segments resulting in complete loss of data (Fig. 2D)[16].

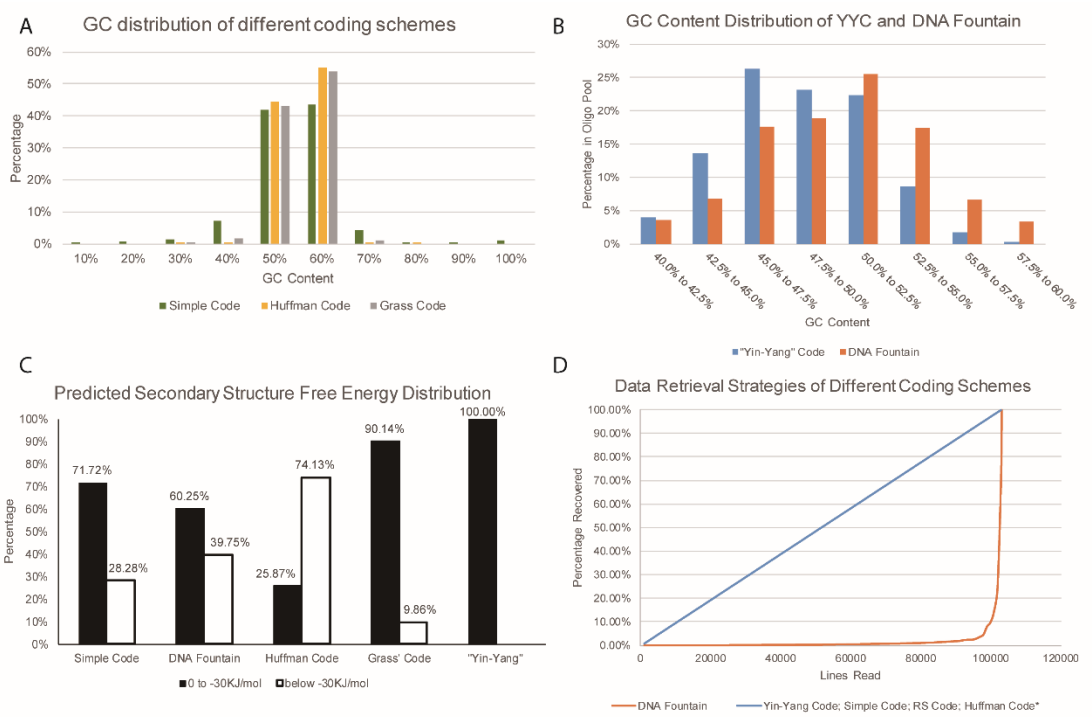


Figure 2. A) Distribution of GC content of generated oligos that transcoded using Simple code (green), Huffman code (yellow) and Grass' code (grey); B) Distribution of GC content of generated oligos that transcoded using YYC (blue) and DNA fountain (orange); C) Percentage of free energy of predicted secondary structure of generated oligos that transcoded using

different codec algorithms, black: free energy between 0 to -30kcal/mol, white: free energy below 30kcal/mol; D) Data retrieval strategies of different coding schemes, DNA fountain (orange) exhibits an exponential growth curve with a long lag-phase, while other coding methods including YYC (blue) suggest a linear growth curve.

Prior to YYC, DNA Fountain is a pioneering DNA-based data storage scheme, which introduces strategies from communication field. It also innovatively employs a screening process to filter sequences with long homopolymer or unbalanced GC content. Traditional fountain code, or Luby Transfrom (LT) code ensure successful decoding by satisfying matrices full rank, i.e. all rows are linearly independent[17, 18]. However, DNA Fountain demands more redundancy other than orthodox LT code for a successful data retrieval, to fulfill the requirements of full rank as well as sequence constraints using screening process.

To prove the robustness and feasibility, more transcoding simulation tests were performed to compare DNA Fountain and YYC. Source code of DNA Fountain was used to perform encoding and decoding tests on 9 different formats of files and 10 bitmap images. As shown in Supplementary Table. S1 and Table. S2, the redundancies needed for different files vary based on file format and data structure, which increase the practical difficulties because the minimum redundancy should be verified for each individual file before transcoding, and also demonstrated variable information density (>1.12 fold of information density of YYC in this experiment). Current solution for this issue is to unify the file format by compression through tar as used in previous works with DNA Fountain[10, 19]. However, the compression risks that the whole original data may be lost once the compressed file is corrupted. Furthermore, we found that files with particular data structure, such as the tricolor flag bitmap images, may cause inability of data retrieval despite being encoded with very high redundancy and long runtime.

On the other hand, YYC working scheme transcodes files in their original format without altering the original code element, which will be more robust for data retrieval. As mentioned before, it retrieves source data proportionally to the number of sequences read, which shows better tolerance when some data is corrupted. Meanwhile, the redundancy is set according to practical requirements and extent of robustness of data needed. In this *in silico* transcoding tests, all the transcoded files exhibit a higher

information density using YYC than using DNA Fountain. Therefore, YYC would be more practical and robust for further applications of DNA-based data storage.

It also should not be neglected that researchers are still trying to further increase the coding potential from other perspectives. In 2019, Choi et. al augmented the information capacity per nucleotide by using degenerate bases[20]. Recently, Anavy et. al reported a strategy by harness the multicopy features of current DNA synthesis technology using composite DNA letters[21]. Both strategies can improve the theoretical coding potential according to theoretical calculation and their proof-of-concept demonstrations. However, both strategies require nucleotides mixtures to be added to oligonucleotides in one column or spot. Meanwhile, in order to differentiate inevitable sequencing errors and predesigned degenerate base or composite letter, they also require much more accurate sequencing result to obtain the distribution of 4 bases and thus more sequencing will be performed. Therefore, these strategies are neither quite practical nor robust under current DNA synthesis and sequencing technologies.

Experiment suggests feasibility of YYC codec system

As a proof of concept, the 'Yin-Yang' codec system was applied to encode three digital files into a synthesized oligo pool, including a 113 kb JPG image showing the mechanism of DNA replication[22], a 97 kb English poem collection from the Shakespearian Sonnets, and a 36 kb ancient Chinese classic - Tao Te Ching (Fig. 3A). The oligos in the synthesized pool possess the same structure, including flanking primer regions (20 nt in both ends), data payload region (120 nt), Hamming code region (8 nt) and index region (14 nt)(Fig. 3B). The transcoding yielded 11098 oligo sequences with 182 nt each, including 33.3% of redundancy in case of sequence loss.

After synthesis, a serial dilution experiment for the oligo pool was performed in order to determine the minimal average copy number of oligos required for file retrieval. PCR amplification was conducted to generate double-stranded DNA libraries for long-term storage and sequencing afterwards. All the PCR amplifications were successful for diluted samples. Moreover, it is shown that by using file-specific primers, we can amplify specific files from the pool[7, 23, 24] (Fig. 3C). The quantities of each file are also of consistent quantity according to gel electrophoresis, which suggested the feasibility of random-access.

The amplified DNA library was then deep-sequenced (average sequencing depth = 4000x) using BGI-Seq-T1 NGS sequencer. For accuracy, reads with length not equal to 182 was first removed and gave approximately half of the original sequencing data because of synthesis errors and barcode contamination during library construction for sequencing. The sequencing data shows that, with sequencing in this instrument, even with high sequencing depth, loss of few oligos is inevitable (Supplementary Fig. S4). This might be caused from many processes, including synthesis, necessary manipulation like PCR amplification, probe capture assay or other molecular approaches, storage, transportation of samples, and even sequencing itself. It can be concluded that to prevent loss of data, a certain amount of redundancy should be added. Errors introduced in DNA synthesis and sequencing were considered to be a major issue in data retrieval. Hence, error correction codes are used to ensure the fidelity. However, it is interesting that, with the sequencing depth increasing, the errors can be corrected by sequencing alignment as well as consensus sequence determination[25]. Therefore, it can be suggested that with sufficient sequencing depth, extra redundancy is necessary for DNA-based data storage, but not error-correction codes.

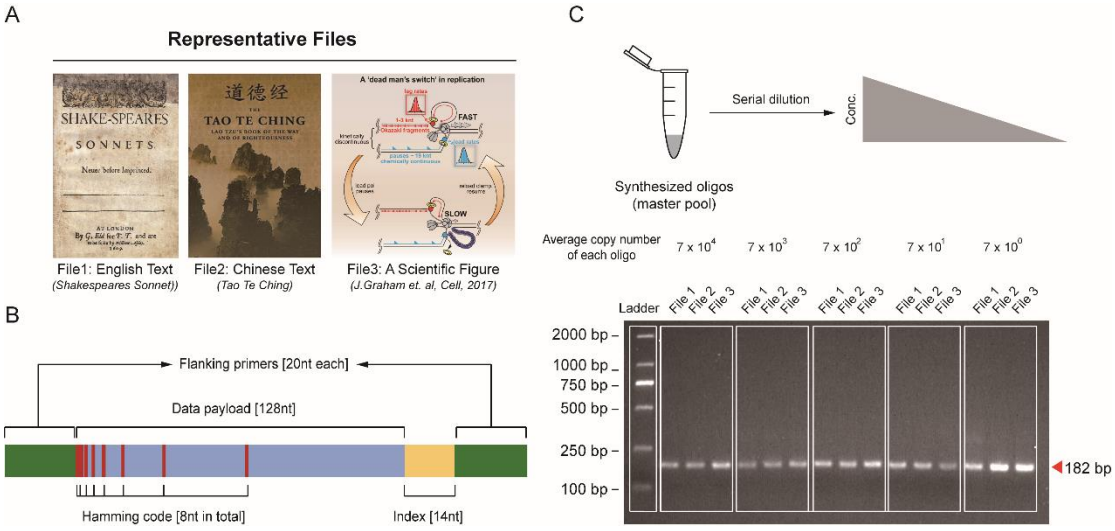


Figure 3 (A) Three files used for DNA-based data storage in this work, including Chinese and English text and image. (B) Structure of oligos. (C) Serial dilution experiment on deep copying of oligo pool which have the original oligo average copy number of $7x10^7$. All the diluted samples can be correctly amplified. Electrophoresis results of PCR amplification products using synthesized oligo pool as template but different primers and also different diluted samples.

Discussion

Apart from the outstanding performance in coding potential and valid DNA sequence generation, YYC also shows broad application not only because of its variant rules but also the flexibility of incorporation. In the current pipeline, the flexible incorporation is used in screening process in order to obtain valid DNA sequences according to preset criteria. Furthermore, unlimited strategies could be used for choosing binary segments to be incorporated (e.g. random choice or choice following predesigned patterns), which give multiple choices for data manipulation and broaden the application scenario. For example, classified information encryption, the incorporation pattern can be various and thus provide a more secured data archiving. On the other hand, although YYC has 1536 derivative rules, different rule exhibits different efficiency in runtime of program of transcoding identical file. It would be a huge workload to determine the optimal rule for particular file format or data structure.

To summarize, in this work, we reported a robust and efficient codec system which displays outstanding feature in coding potential and sequence generation that is more amenable to synthesis and sequencing. We tested our codec algorithm *in silico* with collection of different types of files including text, image, audio, video, etc. These results suggest a better overall performance for data density and robustness for our transcoding strategy compared to those previously reported. A wet experiment also proved the feasibility and also showed the capability of our pipeline for random-access of files. In addition, the experiment demonstrated the highest physical information density to date for successful data retrieval. We expect that this codec system could provide more choices for people to use DNA-based data storage more efficiently with broader applications.

Moving forward, it should not be a major issue to improve the coding potential in current DNA-based data storage since DNA Fountain and YYC are already capable of reaching the theoretical limit. The physical maximum information density of DNA-based data storage could also be improved if researchers could find better strategies to make deep copies from micro-synthesized DNA sequences, which have very small average copy number. Meanwhile, instead of file-level random-access, it is also important to develop other novel techniques in smaller scale. For example, Tomek et. al reported an emulsion-PCR based random-access technique[13], which can also be used

for sub-file level random access if reasonable partition rules and primers are designed. Although the cost and speed of DNA synthesis and sequencing are still far behind silicon-based storage system, DNA-based data storage is currently considered as a potential solution for huge data long-term archiving. And with DNA synthesis and sequencing technology improving, we expect that DNA-based data storage could gradually be cheaper, faster and more secure way of immediate access storage for people's daily lives.

Currently, other innovative molecule storage medium have been proposed including oligopeptides[3, 26] and metabolomes[4], which are used for binary transcoding. Their variety can achieve a greater data compressibility compared to DNA, which have four different kinds of bases. However, the stored information cannot be amplified for information backup, which is a significant requirement for data archiving. Moreover, both techniques use array chips for storage and mass spectroscopy for information retrieval. As a result, the data density could be a problem as well as the widespread application. As both works concluded, these techniques have great space for optimization. Hence, we may suggest that before breakthrough techniques are developed, DNA-based information storage would still be a reliable storage strategy for mass data archiving in the not-so-far future.

Methods

'Yin-Yang' DNA-based data storage pipeline

The 'Yin-Yang' codec pipeline includes three major steps: segmentation, incorporation and validity-screening (Supplementary Fig. S2). (1) Binary information is extracted from source file and then partitioned into smaller segments according to requirements. Indices will be added to each segment and a pool of binary segments with identical length will be obtained. (2) Two segments will be selected randomly or following certain patterns, and then incorporated into one DNA sequence according to 'Yin-Yang' codec algorithm. Considering the features of incorporation algorithm, binary segments contains too much '0' or '1' tends to produce DNA sequences with abnormal GC content or undesired repeats. Therefore, for the convenience to reduce the calculation in this demonstration, binary segments contain high ratio of '0' or '1' (>80%)

will be collected into a separate pool and will be firstly selected to incorporate with binary segments with normal ratio of '0/1'. (3) In order to prevent the difficulties of synthesis and sequencing, the incorporated sequence will be tested through a validity-screening process, including checking their GC content, maximal homopolymer length, secondary structure free energy, etc., and only the sequences meet the criteria will be considered as valid sequences. The details of validity-screening are described in the following section.

YYC-screener to tackle biochemical constraints

Biochemical features of DNA molecules including high GC or AT content, homopolymer and complex secondary structure are usually quite harsh for normal DNA synthesis and sequencing procedure, therefore results in difficulties of writing and reading process in DNA-based data storage applications. It has been reported that DNA sequence with over 6-mer homopolymers or a stable secondary structure (of low free energy) will cause many problems for sequencing[27, 28]. Additionally, DNA sequences with GC content below 40% or above 60% often bring troubles to DNA synthesis manufacturers. For practical DNA-based data storage usage, it is therefore important to build a working scheme to reduce these situations. By harnessing the advantage of assembling two binary information to one DNA sequence in 'Yin-Yang' codec system, a working scheme named 'YYC-screener' is established here to obtain the valid sequences. As mentioned previously, two sequences with different complexity are selected and incorporated according to YYC. Generated sequences with GC content over 60% or less than 40%, or carrying >6-mer homopolymer, or possessing a predicted secondary structure of <-30 kcal/mol are abandoned and the two original segments are put back into the pool. A new pair of segments will then be selected to repeat the screening process until it passes the validity-screening test (Supplementary Fig. S2). We performed this working scheme on ~1GB data including articles, images, audios and videos. It is found that YYC-screener can generate DNA sequences which meet the criteria as expected.

File encoding

The 3 files' binary form (9.26×10^5 bits, 7.95×10^5 bits and 2.95×10^5 bits) were extracted respectively and segmented into three 120 bit-segment pools. First, we added 8-bit Hamming code for error correction, which is able to correct at most two mutation

errors possibly caused by synthesis and sequencing. After this, three 128-bit binary segments (data payload + Hamming code) were used to generate a fourth redundant binary segment to prevent possible sequence loss during the experiment. Furthermore, 14-bit index information was added into each binary segment to infer their address in the digital file and in the oligo mixture for purpose of decoding. Rule 888 from 'Ying-Yang' coding scheme was then used to convert the binary information into DNA bases. To reduce the sequence complexity, the aforementioned 'YYC-screener' was used to screen the transcoded DNA sequence. Sequences passed the screening all satisfied the criteria of balanced GC content (40% to 60%), higher secondary structure free energy (> -30kcal/mol) and minimum (< 6) homopolymer repeats. These enabled us to generate 8323 DNA sequence segments with 142 nt each. 33.3% redundancy was added in case of possible sequence loss during manipulation through exclusive-or calculation[12]. For the future reading and random-accessing of each file, well-designed flanking sequence of 20 nt was added at both ends of each DNA segment. Finally, an oligo pool containing 11098 single-stranded DNA sequences of 182 nt was sent to the DNA supplier for synthesis (Fig. 3B).

Polymerase Chain Reaction (PCR) amplification of synthesized DNA oligo pool

Before NGS sequencing, the single-stranded DNA oligo pools were amplified by PCR reactions. Five independent reactions were conducted to obtain the amplified product of three different files respectively. Q5 polymerase was used in these reactions in order to ensure the fidelity and 19/27/32/35/39 cycles were set to verify the minimum average copy number of oligos for data retrieval. The experiment conditions are shown below.

Table 2. Solution component for PCR amplification.

Component	Volume（μL）
5 × Q5 Reaction Buffer	10
Q5 polymerase	0.5
dNTPs（10mM）	4
Forward Primer（10μM）	2
Reverse Primer（10μM）	2
Template DNA （×100/1000/10000）	1
ddH_2O	30.5

Table 3. Experiment condition for PCR amplification.

Reactions	Temperatures	Time	Cycles
Pre-denaturation	98℃	5min	×1
Denaturation	98℃	15s	
Annealing	62℃	30s	×19/27/32/35/39
Synthesis	72℃	10s	
Extension	72℃	5min	×1

Conflict of Interest

Sha Zhu is a currently an employee at the Sensyn Health Plc. This work was completed when Sha Zhu was working at the University of Oxford, and consulting for the BGI. George Church significant interests in Twist, Roswell, BGI, v.ht/PHNc, v.ht/moVD.

Acknowledgements

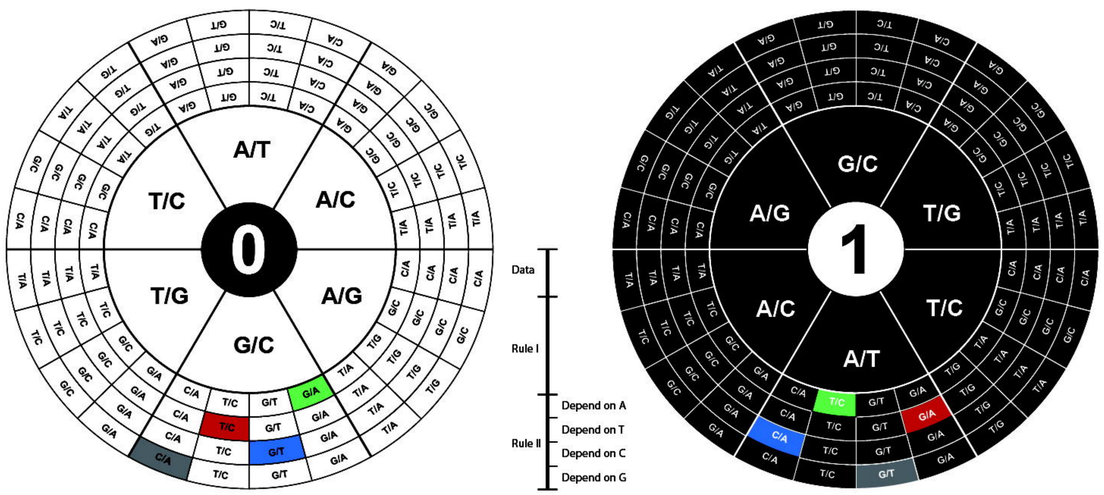
This work was supported by the Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (No. 2017B090904014) and Guangdong Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011). This work was also supported by the George Church Institute of Regenesis, BGI-Shenzhen, China.

Note: The data that support the findings of this study have been deposited in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb with accession code CNP0000748.

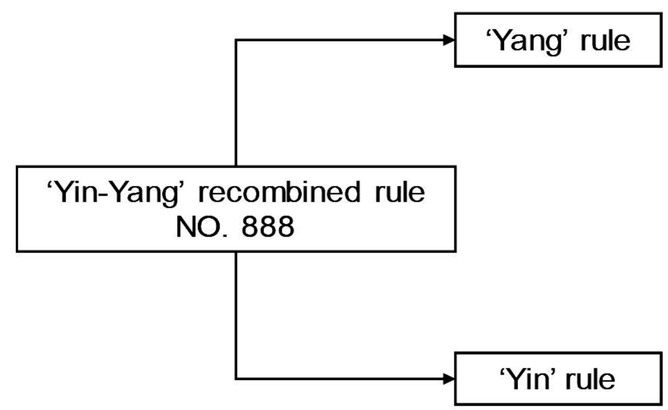
Reference

1. Church, G.M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337**, 1628 (2012).
2. Allentoft, M.E. et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc Biol Sci* **279**, 4724-4733 (2012).
3. Cafferty, B.J. et al. Storage of Information Using Small Organic Molecules. *ACS Cent Sci* **5**, 911-916 (2019).
4. Kennedy, E. et al. Encoding information in synthetic metabolomes. *PLoS One* **14**, e0217364 (2019).
5. Kosuri, S. & Church, G.M. Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods* **11**, 499-507 (2014).
6. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature* **550**, 345-353 (2017).
7. Yazdi, S.M., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A Rewritable, Random-Access DNA-Based Storage System. *Sci Rep* **5**, 14138 (2015).
8. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77-80 (2013).
9. Grass, R.N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W.J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew Chem Int Ed Engl* **54**, 2552-2555 (2015).
10. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950-954 (2017).
11. Ping, Z. et al. Carbon-based archiving: current progress and future prospects of DNA-based data storage. *Gigascience* **8** (2019).
12. Organick, L. et al. Random access in large-scale DNA data storage. *Nat Biotechnol* **36**, 242-248 (2018).
13. Tomek, K.J. et al. Driving the Scalability of DNA-Based Information Storage Systems. *ACS Synth Biol* **8**, 1241-1248 (2019).
14. Zhirnov, V., Zadegan, R.M., Sandhu, G.S., Church, G.M. & Hughes, W.L. Nucleic acid memory. *Nat Mater* **15**, 366-370 (2016).
15. Bornholt, J. et al. in Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems 637-649 (2016).
16. Chen, K., Zhu, J., Boskovic, F. & Keyser, U.F. Secure data storage on DNA hard drives. *bioRxiv*, 857748 (2019).
17. Huang, W., Li, H. & Dill, J. in 2011 Wireless Telecommunications Symposium (WTS) 1-5 (2011).
18. Asteris, M. & Dimakis, A.G. Repairable Fountain Codes. *Ieee J Sel Area Comm* **32**, 1037-1047 (2014).
19. Koch, J. et al. A DNA-of-things storage architecture to create materials with embedded memory. *Nat Biotechnol* **38**, 39-43 (2020).
20. Choi, Y. et al. High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Sci Rep* **9**, 6582 (2019).
21. Anavy, L., Vaknin, I., Atar, O., Amit, R. & Yakhini, Z. Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nat Biotechnol* **37**, 1229-1236 (2019).
22. Graham, J.E., Marians, K.J. & Kowalczykowski, S.C. Independent and Stochastic Action of DNA Polymerases in the Replisome. *Cell* **169**, 1201-1213 e1217 (2017).
23. Kosuri, S. et al. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol* **28**, 1295-1299 (2010).
24. Tian, J. et al. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050-1054 (2004).
25. Lee, H.H., Kalhor, R., Goela, N., Bolot, J. & Church, G.M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nat Commun* **10**, 2383 (2019).
26. Cafferty, B.J. et al. Storage of Information Using Small Organic Molecules. *ACS Central Science* **5**, 911-916 (2019).
27. Kulski, J.K. Next-generation sequencing—an overview of the history, tools, and "Omic" applications. *Next Generation Sequencing–Advances, Applications and Challenges*, 3-60 (2016).
28. Kieleczawa, J. Fundamentals of sequencing of difficult templates--an overview. *J Biomol Tech* **17**, 207-217 (2006).

A



B



Local Nucleotide Candidates

Binary Digits

0	1
A or T	C or G

Binary Digits

Previous Nucleotide	0	1
A	A or G	C or T
T	C or T	A or G
G	A or G	C or T
C	C or T	A or G

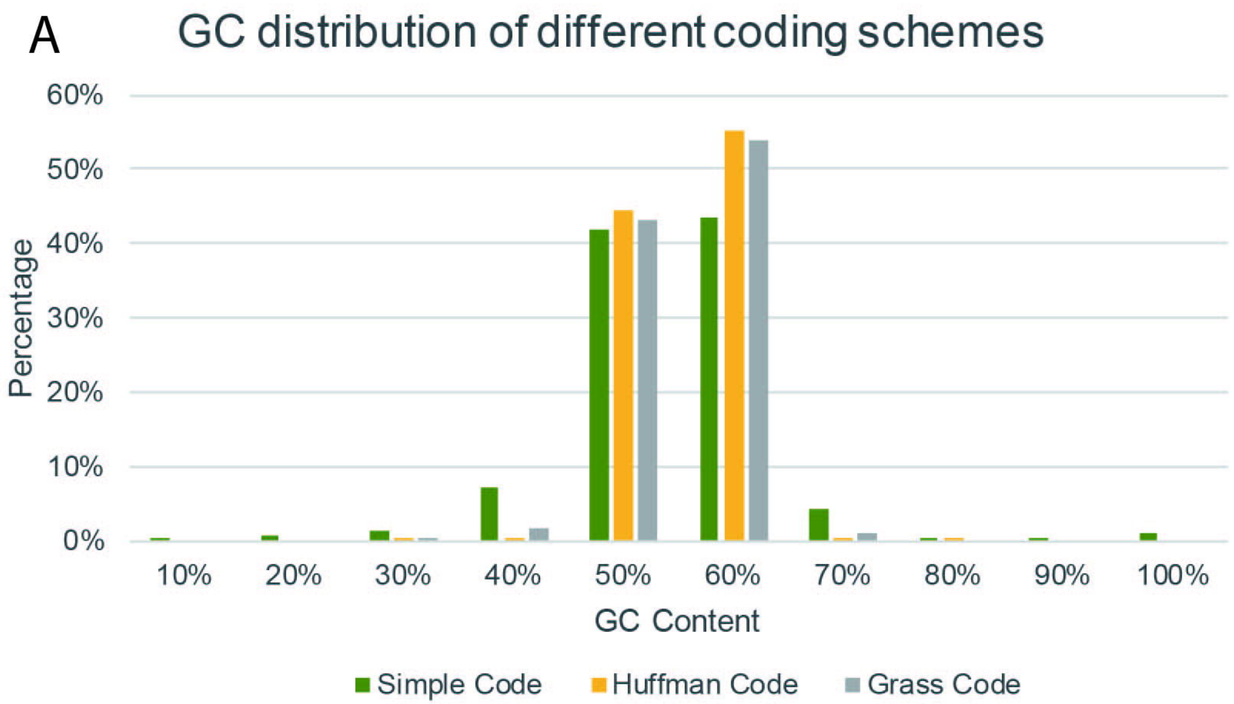
C A step-by-step transcoding demonstration using rule No. 888

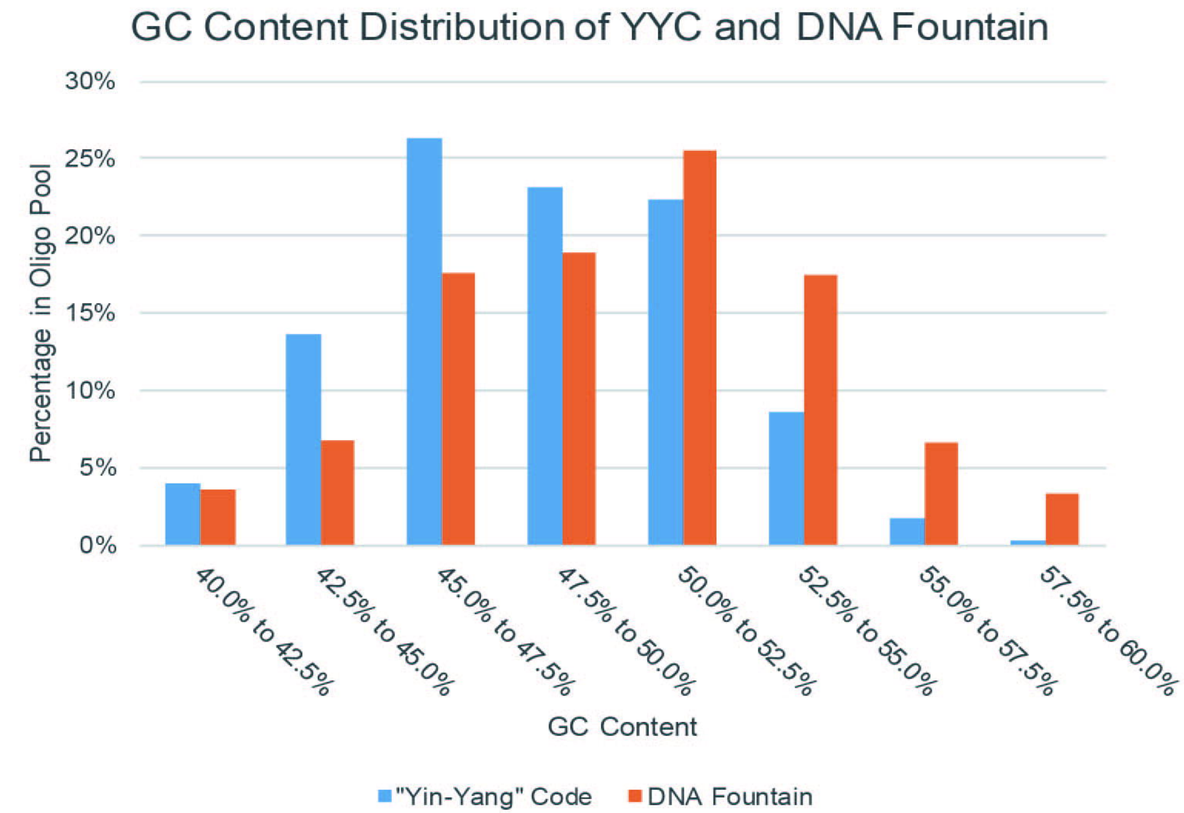
Input Signals a = 10110011 b = 01011101

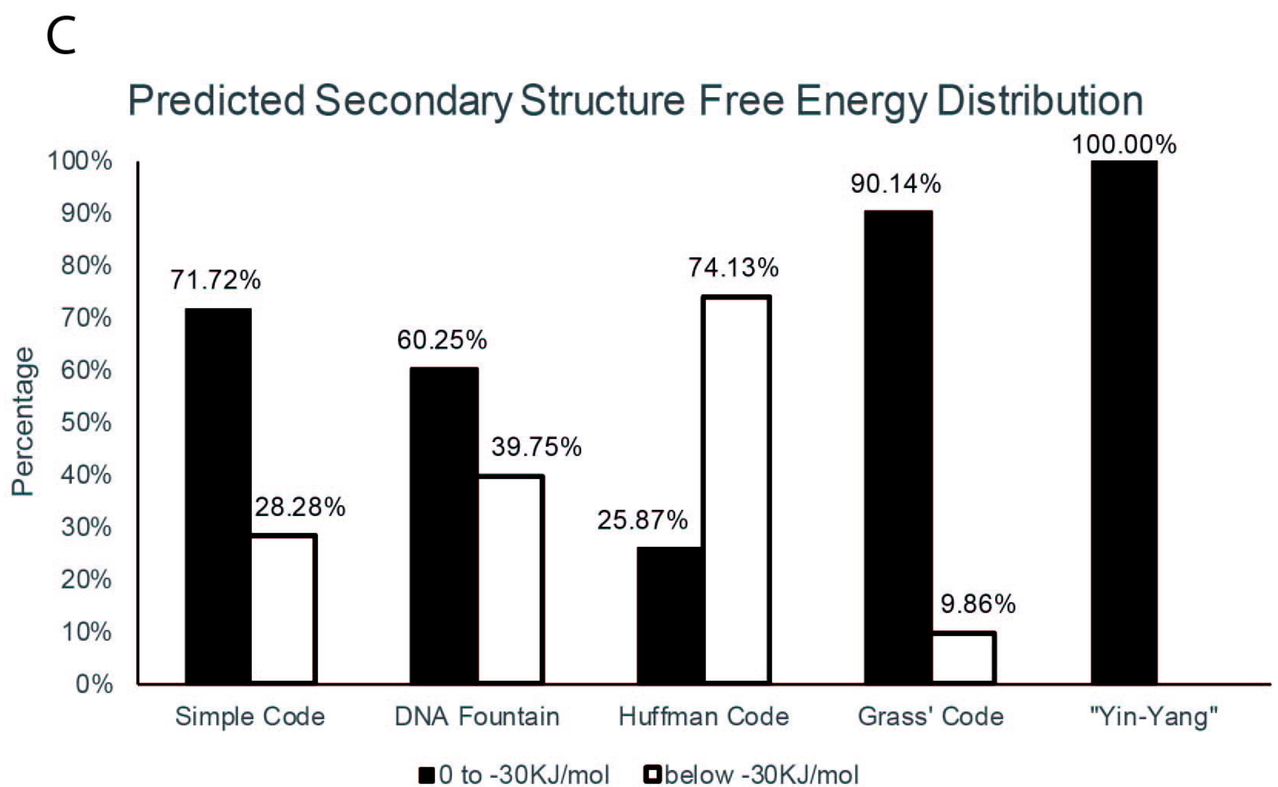
Transcoding According to Rules

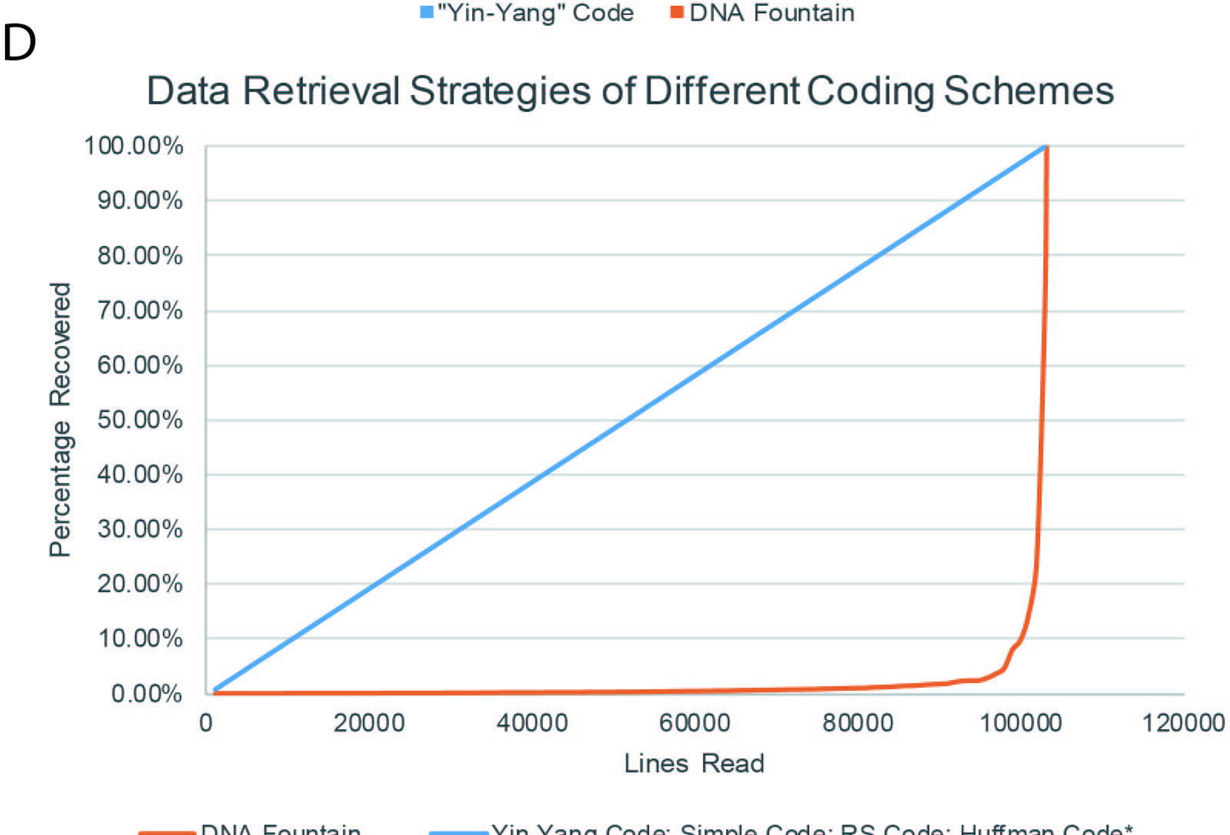
	Previous Nucleotide	Binary Digits	Rule	Candidates	Local Nucleotide	Oligo Sequence
Step 1	A virtual	a1 = 1 b1 = 0	'Yang' 'Yin'	C or G A or G	G	G
Step 2	G	a2 = 0 b2 = 1	'Yang' 'Yin'	A or T C or T	T	GT
			. . .			
Step 8	C	a8 = 1 b8 = 1	'Yang' 'Yin'	C or G C or T	C	GTCGTAGC

Result Sequence: GTCGTAGC


A
GC distribution of different coding schemes
Percentage
60%
50%
40%
30%
20%
10%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
GC Content
Simple Code
Huffman Code
Grass Code

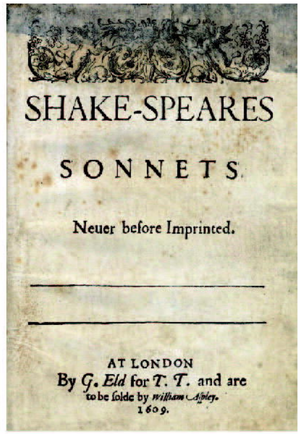

B
GC Content Distribution of YYC and DNA Fountain
Percentage in Oligo Pool
30%
25%
20%
15%
10%
5%
0%
40.0% to 42.5%
42.5% to 45.0%
45.0% to 47.5%
47.5% to 50.0%
50.0% to 52.5%
52.5% to 55.0%
55.0% to 57.5%
57.5% to 60.0%
GC Content
"Yin-Yang" Code
DNA Fountain


C
Predicted Secondary Structure Free Energy Distribution
Percentage
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
71.72%
28.28%
60.25%
39.75%
25.87%
74.13%
90.14%
9.86%
100.00%
Simple Code
DNA Fountain
Huffman Code
Grass' Code
"Yin-Yang"
0 to -30KJ/mol
below -30KJ/mol

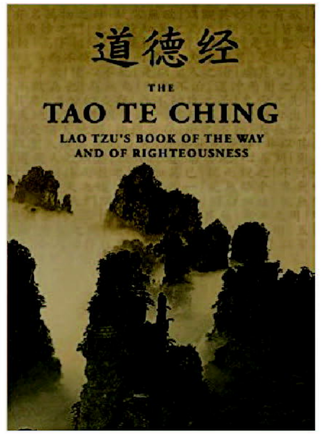

D
Data Retrieval Strategies of Different Coding Schemes
Percentage Recovered
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
20000
40000
60000
80000
100000
120000
Lines Read
DNA Fountain
Yin-Yang Code; Simple Code; RS Code; Huffman Code*

A

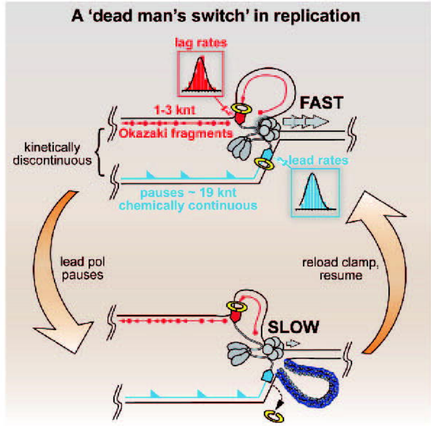
Representative Files



File1: English Text
(Shakespeares Sonnet))

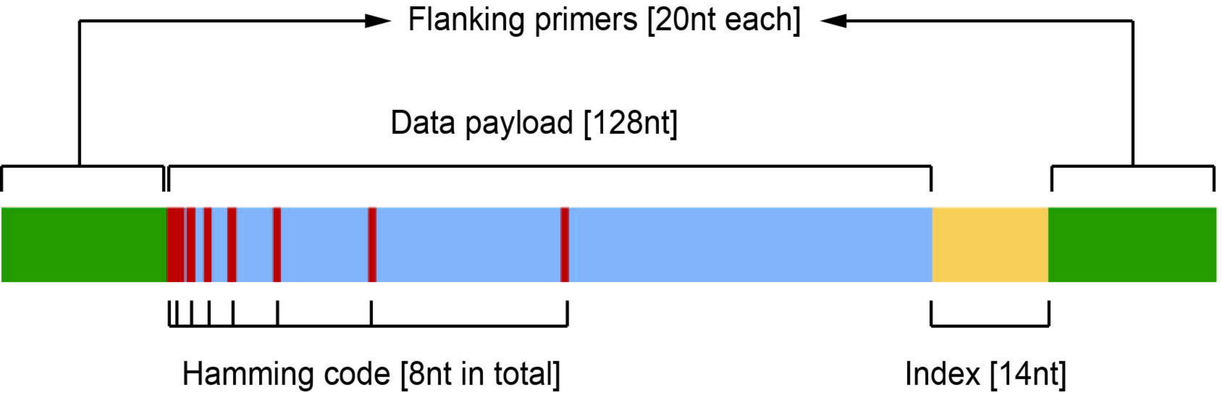


File2: Chinese Text
(Tao Te Ching)

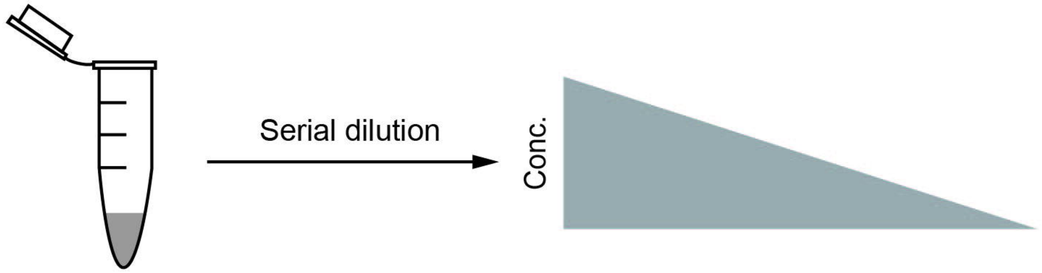


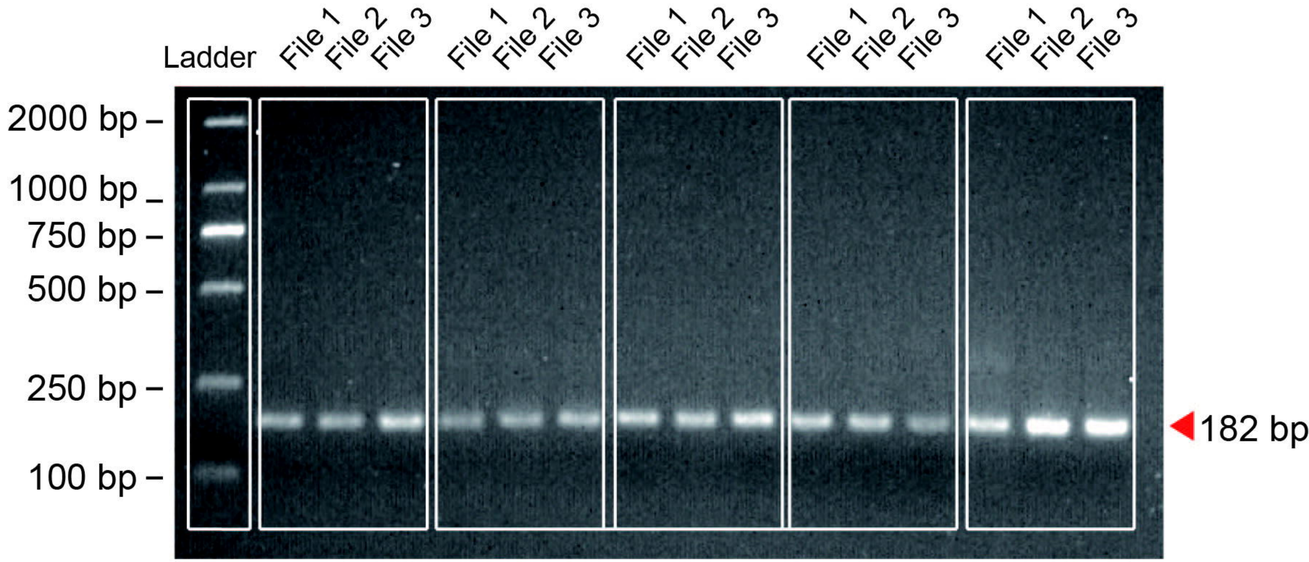
File3: A Scientific Figure
(J.Graham et. al, Cell, 2017)

B



C



Synthesized oligos
(master pool)

Average copy number of each oligo	7×10^4	7×10^3	7×10^2	7×10^1	7×10^0