Materials Today Bio 28 (2024) 101221

Contents lists available at ScienceDirect

Materials Today Bio

journal homepage: www.journals.elsevier.com/materials-today-bio

"Multi-layer" encryption of medical data in DNA for highly-secure storage

Jiaxin Xu [a,b,1], Yu Wang [b,1], Xue Chen [b], Lingwei Wang [a], Haibo Zhou [c], Hui Mei [b,**], Shanze Chen [a,*], Xiaoluo Huang [b,***]

[a] Department of Pulmonary and Critical Care Medicine, Institute of Respiratory Diseases, Post-doctoral Scientific Research Station of Basic Medicine, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University, The First Affiliated Hospital of Southern University of Science and Technology), Shenzhen, 518020, Guangdong, China

[b] Shenzhen Key Laboratory of Synthetic Genomics, Guangdong Provincial Key Laboratory of Synthetic Genomics, Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, Guangdong, China

[c] College of Pharmacy, Jinan University, Guangzhou, Guangdong, 510632, China

ARTICLE INFO

Keywords:
Medical data
Data security
DNA data storage
"Multi–layer" encryption
Mass spectroscopy

ABSTRACT

The exponential increasement and the attributes of medical data drive the requirement for secure medical data archiving. DNA data storage shows promise for storing sensitive and important data like medical records due to its high density and endurance. Nevertheless, current DNA data storage working scheme generally does not fully consider the data encryption, posing a risk of data corruption by routine DNA sequencing. Here, we designed a "multi-layer" encryption pipeline for medical data archiving. Initially, digital information was encrypted using Blowfish algorithm at information technology (IT) layer, followed by two-layer data encryption at the biotechnology (BT) layer. The first BT layer exploited the molecular weight of synthetic DNA or nucleoside to encrypt the key, while the second BT layer encrypted digital information within DNA sequences. Consequently, decryption involved layer-by-layer interpretation of data, including mass spectroscopy, sequencing, and Blowfish decryption, significantly enhancing data security. Utilizing mass spectroscopy to retrieve information allows for employment of both natural and unnatural nucleosides, as well as their synthetic oligonucleotides, for data storage, thereby considerably boosting scalability. Our work implies expanded flexibility of DNA-based data storage, highlighting the potential for leveraging various physical and chemical characteristics of DNA molecules to encode and access digital information.

1. Introduction

Ever since the computer revolution, unfathomable data has been produced, with an estimation of data reaching 175 Zettabyte in 2025 [1]. Among diverse branches of data, healthcare data is projected to experience a faster annual growth of 36 % than other industries such as financial services (26 %), manufacturing (30 %), and media and entertainment (25 %) by 2025 [2–4]. Currently, hospitals are troubled by the long-term and private storage of medical data. The risk of medical data leakage poses a significant threat, potentially resulting in serious consequences such as compromising patient safety, damaging reputation, as well as causing financial losses, which ultimately affects doctor-patient relationships. Therefore, it is critical to look for highly-secure and long-term manner for medical data storage.

DNA has recently been regarded as a promising medium for data storage owing to its immutable nature, high latency, and unprecedented data density (455 exabytes per gram of ssDNA) [5–8]. It offers 10^7 folds higher information density than conventional storage devices, with one single gram of DNA storing 40 zettabytes data, while 277 million of magnetic tapes or 42 billion of USB sticks are required to store the same amount of data [9,10]. In addition, it has been reported that DNA has a 10^5 longer retention time compared to flash memory [11]. Owing to its high density and long endurance, DNA shows its great potential for highly secure data archiving of the medical data.

Standard DNA-based data storage scheme includes encoding, synthesis, storage, sequencing, and decoding [12,13]. Significant

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: hui.mei@siat.ac.cn (H. Mei), chenshanze@mail.sustech.edu.cn (S. Chen), huangxl@siat.ac.cn (X. Huang).
[1] These authors contributed equally.

https://doi.org/10.1016/j.mtbio.2024.101221
Received 8 July 2024; Received in revised form 21 August 2024; Accepted 31 August 2024
Available online 1 September 2024

breakthroughs in DNA synthesis and sequencing technologies fueled the rapid development of DNA-based information storage since 2012, with 659 KB and 739 KB of data being stored in DNA molecules [5,14]. Efforts have not only focused on enlarging the amount of information stored in DNA, but also on improving reliability and storage density. For instance, in 2015, Grass et al. proposed a platform that achieved stable chemical storage by encapsulating information-encoded DNA in silica [15]. Moreover, 'DNA fountain code' was proposed to offer high data storage density, leaving profound impact on the subsequent works such as the 'DNA-of-things' (DoT) that explores the scalability of storing DNA information in various materials [16,17]. However, current DNA data storage working schemes generally composed of natural DNA nucleosides [5,18]. This might bring drawbacks for data encryption, as the data may be invaded by ordinary DNA sequencing [19]. Therefore, exploring innovative architectures to avoid direct decoding information from DNA sequence should be necessary to address these challenges. While DNA sequence information has been widely used for data storage, the molecular weight of "DNA" or "nucleoside" for data storage has been never explored, therefore leaving a possibility for encryption of data out of "sequence" level.

Herein, we report a novel "multi-layer" encryption architecture for medical data storage, essentially by the integration of "molecular weight" based mass spectroscopy (MS) DNA storage. In specific, an algorithm was firstly built to store the key, "BELIEF", into "molecular weight" of "nucleosides" and synthetic DNA. By doing this, both natural and unnatural bases can be used for this storage. The density and capacity of MS-based data storage are simulated, displaying distinct performance compared to traditional DNA data storage. Based on this, a "three-layered" data protecting strategy was established to archive the medical report, with the IT layer achieved by Blowfish encryption algorithm to generate a key (first layer), the two BT layers encode the key in molecular weight (second layer) of either oligonucleotides or nucleosides and the report in DNA sequence contents (third layer). Accordingly, data retrieval relies on the integration of MS and sequencing, followed by deciphering using Blowfish decryption algorithm. Metal-assisted laser desorption/ionization (MALDI) and liquid chromatography-mass spectroscopy (LC-MS) are employed to read one portion of the data stored in DNA oligonucleotide and nucleosides, respectively (Fig. 1). The proposed "multi-layer" encryption for DNA data storage illuminates the information storage and reading at diverse level, paving the way for utilizing various IT algorithms and various physical and chemical properties of molecules for data storage, such as structure, length, UV or infrared absorbance, conductivity, *etc.*

2. Materials and methods

2.1. Encryption and encoding strategy

The encryption and encoding are implemented to achieve secure data storage (Table S1). To ensure compatibility between the length of the key and the MS information encoding, we utilize the symmetric encryption algorithm Blowfish to encrypt the original information. The Blowfish is a classical, symmetric encryption algorithm, which is explained in detail in Supplementary note 1 [20–22]. Initially, a cipher is constructed using the CBC mode and a specified "key". Subsequently, the cipher encrypts the information after being padded. Following this encryption step, we employ Storage-D, an intuitive online platform, for encoding the encrypted information [23]. This platform offers a range of classic encoding algorithms as well as a proprietary Wukong encoding method (with a potential density of approximately 1.98 bit/nt). The Wukong encoding algorithm allows customization of DNA chain length, GC content, as well as incorporating error correction codes of any lengths, *etc*. In this study, we utilize the Wukong encoding algorithm to encode encrypted information into DNA with default parameters (Encoded Length = 1400, Homopolymer = 4, Max GC = 40 %, Min GC = 60 %, ECC = 4).

2.2. Mass spectroscopy

The oligonucleotides were dissolved in nuclease-free water to 100 μM, while all nucleoside monomers were dissolved in CH_3OH with a concentration lower than 1 mg/mL. MALDI was performed to obtain the mass information of oligonucleotides using Autoflex Max LRF (Bruker). The matrix preparation involved dissolving Ammonium citrate dibasic

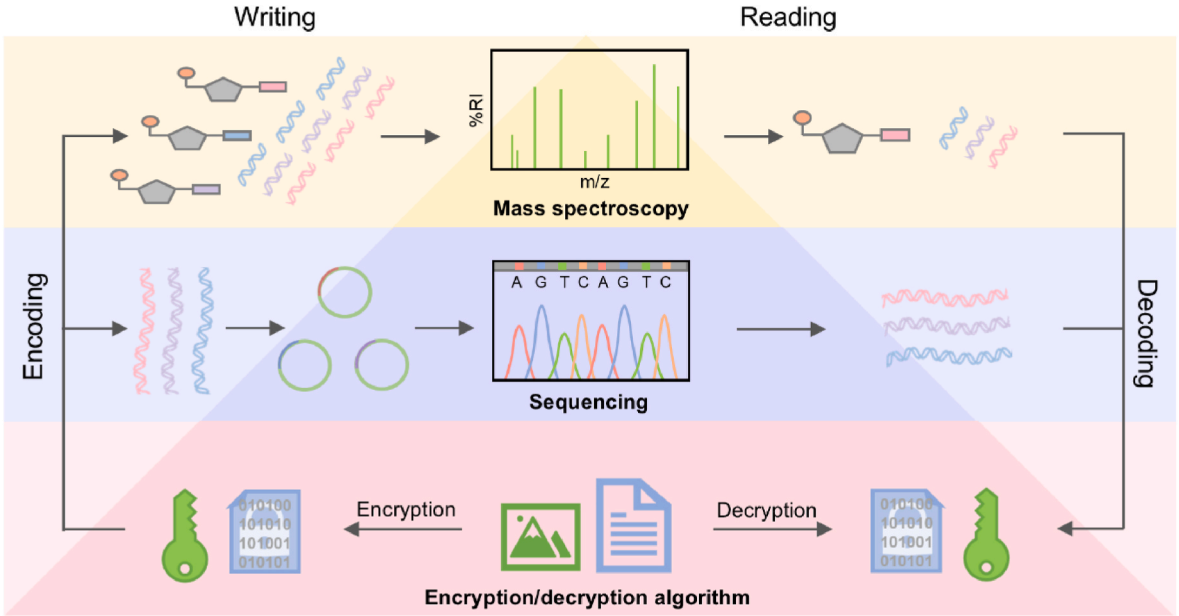

Fig. 1. Illustration of "multi-layer" encryption for DNA data storage. "Multi-layer" refers to the encryption of digital information through a combination of information technology (IT) and biotechnology (BT) strategies. Text or image files can be encrypted using the currently available algorithm, the symmetric-key block Blowfish cipher (first layer, IT layer). The "key" is then encoded by the molecular weights of either nucleosides or DNA oligonucleotides, while the encrypted information is encoded in DNA sequences. There are regarded as the second layer and the third layer. (BT layers). To decode the information accurately, two-layer characterization is performed through an integration of MS and DNA sequencing, followed by one-layer decipher using Blowfish decryption algorithm.

(DCA) in 50 % (v/v) of water in acetonitrile (MeCN) to prepare 10 mg/mL DCA solution, which was used as solvent for 3-Hydroxypropionic acid (3-HPA) to get a saturated solution of 3-HPA. Subsequently, 0.5 μL of the matrix solution was spotted onto the sample target for drying, followed by spotting equal amounts of samples onto the dried matrix, which were then dried again. The spectra were acquired in linear positive mode, and data analysis was performed by mMass. To read the mass information of nucleosides, LC-MS analysis was performed using UltiMate 3000 (Thermo Fisher Scientific) coupled with Q Exactive (Thermo Fisher Scientific). The mobile phase consisted of 0.1 % formic acid and 4 % MeCN aqueous solutions. Liquid chromatographic separation was accomplished with a Hypersil GOLDTM C18 column (2.1 × 100 mm, 1.9 μm, Thermo Fisher Scientific) with a flow rate of 300 μL/min. Analysis was conducted in ESI + mode with a resolving power of 70000. The ion spray voltage was set at 3500 V, capillary temperature at 320 °C, and orbitrap detection with AGC target ion value of 3 × e6. Data analysis was performed using Xcalbular.

2.3. NMR spectroscopy

All the chemicals were purchased and utilized either before or after accelerated aging without undergoing further purification. The nucleosides were dissolved in dimethyl sulfoxide (DMSO, Innochem). ^{1}H NMR spectra were recorded using a two-channel 400 MHz NMR spectrometer (Bruker).

2.4. Accelerated aging test

The accelerated aging test involved exposing the plasmid in open Eppendorf tubes to different temperatures (60 °C, 65 °C). The temperature control was achieved by a water bath (DK-3, Changzhou Yukun Instrument Manufacturing Co., Ltd), with the samples were suspended in a glass beaker containing a material for humidity control. Saturated sodium bromide (NaBr≥99.5 %, Sigma) solution was utilized to maintain a relative humidity of 50 %. Equal amount of plasmid (200 ng) per tube were subjected to the aforementioned accelerated aging conditions for durations of 1, 2, 3, 5, 7, 9, and 11 days. The sample were either stored at −20 °C or immediately subjected to qPCR.

2.5. PCR and qPCR

PCR was conducted for the plasmid template with a series of dilutions ranging from 10^{1} to 10^{-12} ng/μL using the Biometra TRIO 48 Multi Block thermal cycler (Analytik Jena). For each reaction, 1 μL of plasmid at each concentration was mixed with 1 μL of Ftw primer (10 μM), 1 μL of Rtw primer (10 μM), 2 μL of dNTPs (0.2 mM of each), 0.1 μL of 5 U/μL DreamTaq HotStart DNA polymerase, and 2 μL of 10 × DreamTaq Buffer. The mixture was brought up to 20 μL by nuclease-free water and subjected to an initial denaturation at 95 °C for 1 min, followed by 40 cycles of denaturation at 95 °C for 1 min, annealing at 58 °C for 1 min and extension at 72 °C for 1 min, with a final extension step at 72 °C for 5 min qPCR was conducted for the plasmid before and after accelerated aging using a real-time PCR system (Bio-Rad CFX Duet). Each 20 μL reaction mixture contained an equal amount of plasmid, 0.2 μM of primer 1 and primer 2, 1 × Kod Sybr qPCR mix, 1 × ROX reference dye. The qPCR procedure involved an initial denaturation at 98 °C for 2 min, followed by 40 cycles of denaturation at 98 °C for 10 s, annealing at 60 °C for 10 s and extension at 68 °C for 90 s. Data collection was set during the extension step. Primers used are listed in Table S2.

2.6. Gel electrophoresis

To prepare a 1 % agarose gel, 1g of agarose powder was dissolved into 1 × TAE buffer. Additionally, 10 μL of GelRed Nucleic acid gel stain was added to the gel solution. The gel solution was heated in a microwave oven until it became clear and transparent. After casting, the gel was allowed to polymerize, and the casted gels were stored at 4 °C until use. Prior to loading the samples into each well, samples were mixed with 6 × DNA gel loading dye (Thermo Fisher Scientific). Agarose gel electrophoresis was then performed at a constant voltage of 150 V for 30 min.

3. Results

3.1. The codec principle of molecular-weight based MS DNA data storage

The codec method determines the success of DNA data storage by building a bridge between computer data and DNA molecules. In order to achieve the molecular-weight based DNA data storage, we develop a specialized codec system. The encoding method using molecular weights of nucleosides and oligonucleotides is illustrated in Fig. 2, relying on that the exponential growth in the types of DNA molecules following the increase in DNA sequence lengths. In specific, the numbers of molecules with single nucleoside, two nucleosides, and m nucleosides, are 4, 4^{2}, and 4^{m}, of which the encoded binary bits are 2, 4, and 2m, respectively. Since MS-based molecules identification depends on their molecular weights, two criteria are employed to filter out the molecules with equal molecular weights. On the one hand, $S_i! = S_j\{set(S_i)! = set(S_j)\}$ is applied to eliminate oligonucleotides with the same nucleosides compositions, specifically, oligonucleotide S_i is cleaned out if it contains the same type and number of nucleosides as oligonucleotide S_j, while the different ones are remained. On the other hand, $Wi! = Wj$ is utilized to sweep away oligonucleotides with the same molecular weights (W). Even though large number of molecules are removed, the remained molecules are plentiful for further applications in MS-based data storage. Briefly, the peak positions of the remained molecules are marked on the mass spectra for coding, and all the coding marks are processed for binary encoding. A mass spectrum, as a consequence, will contain a wealth of information. Noteworthily, the storage density for MS-based DNA data storage can be further expanded by the addition of unnatural nucleosides.

As an example, the encoding process to generate the "key" is presented in Figs. S1a and b and lists of DNA sequences and nucleosides are provided in Table S3 and Table S4. Specifically, each letter of the English alphabet is numbered from 1 to 26, with "1" corresponding to "A", "2" to "B", up to "26" for "Z". Consequently, "BELIEF" is translated to "2, 5, 12, 9, 5, 6". The presence or absence of five distinct oligonucleotide chains in mass spectra are designed to represent "1" and "0" in the binary system. Therefore, 5 sequences can be used to represent 2^{5} in binary notation system, with all of the 26 letters included. Based on the above, "BELIEF" is further written as "00010 00101 01100 01001 00101 00110". All in all, "B" is encoded as oligo 4, "E" is represented by a mixture of oligo 3 and 5, and other letters also followed the coding rule in the same way. Alternatively, the "key" can also be stored by nucleosides. In this case, 26 nucleosides are utilized to display the 26 English letters, comprising 4 natural nucleosides and 22 unnatural nucleosides with distinguishable molecular weights.

As illustrated in Fig. S1c, the decoding process to obtain the "key" is performed by detecting the mass spectra of oligonucleotides or nucleosides with information encoded. MALDI is employed to recover and decode the "key" information stored in DNA oligonucleotides, while LC-MS is utilized to read information encoded in nucleosides. The spectra of samples are transformed into binary code, which is then subjected to further decoding operations to extract alphanumeric characters. Inputting the binary code results in the successful decoding of the "key" for information decipherment.

3.2. Performance of molecular-weight based MS DNA data storage

Currently, DNA information storage is still limited by the requirement that data must be written (synthesis) before it can be read (sequencing). Furthermore, DNA synthesis still primarily employs

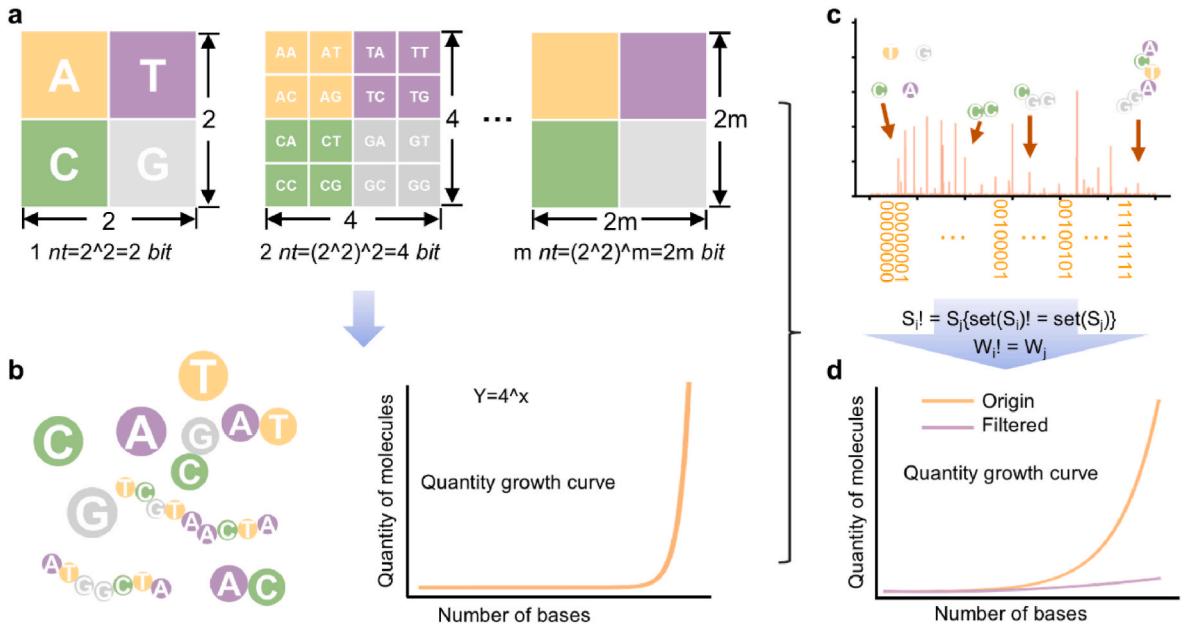

Fig. 2. Principle of nucleosides and oligonucleotides encoding for MS-based DNA data storage. (a) Numbers of molecules made up from 1 nucleoside, 2 nucleosides, until m nucleosides, and the corresponding binary bits. (b) Examples of molecules comprised of various type and numbers of nucleosides, and the exponential increase in quantity of molecules with the number of bases following the equation of $Y = 4^X$, where Y is the quantity of molecules, X stands for the number of nucleosides. (c) Molecules formed by various numbers of nucleosides can be detected by Mass spectrometry, peaks of molecules in mass spectrum are endowed with specific binary strings, which is finally utilized for encoding. (d) Molecules with molecular weight differences smaller than 0.1 *m/z* are filtered out following the equations to obtain easily distinguishable signals by mass spectroscopy, where S and W represent for oligonucleotide and molecular weight, respectively.

natural nucleosides. Alternatively, the developed encoding method relies on scanning mass spectra of bases or oligonucleotide. The approach stores information through single-base encoding with reducing the requirement for long DNA sequences synthesis. Furthermore, MS-based encoding method measures single base or oligonucleotides, and the diversity of oligonucleotide increases with chain length, which produces increasingly more peaks on the mass spectra. As a consequence, this method produces higher storage capacity, not to mention the increasing in density and capacity assisted by unnatural nucleosides addition.

In order to analyze the potential of molecular weight-based MS DNA data storage, we simulated its process as illustrated in Fig. 3a. Unnatural nucleosides are randomly selected from library 1 containing 60 unnatural nucleosides and combined with 4 natural nucleosides to create library 2 (Table S5). A specific number of nucleosides are then selected to form a nucleic acid chain or the so-called codon in DNA-based data storage, resulting in the generation of chain library 3. The molecular weight of each chain is calculated according to the equation $W = \sum_{i=1}^{L} w_i, i \in [1, 64]$. A subsequent filtering operation is applied to eliminate potential chains with identical molecular weight, facilitating easy discrimination by mass spectroscopy. If the mass difference is over 0.1 *m/z*, it contributes to the count value, which is ultimately be utilized for capacity and density calculations.

In the field of DNA digital data storage, information storage density refers to the bit number that a single nucleoside can encode. Reading of information relies on DNA sequencing. Traditional DNA sequencing is independent of the PCR amplification of DNA sequences, which only read the spectral peaks of fluorophores carried by A/T/C/G. Otherwise, sequencing also performed depending on the single-base potential difference of synthetic DNA sequences. Although molecular weight-based MS DNA data storage is quite different from sequence-based DNA data storage, we calculated information density using the similar logic, hopefully giving a reference for the comparison of this method with other methods at the same "concept" level. Briefly, the coding methods for DNA information storage generally use code table mapping, such as Church's work that directly mapped a single base to 0 or 1 [5], Blawat's work mapped A to 00 [24], and the multi code table for yin-yang codes [25]. In our work, the 1, 2, and n codons represent for the codon comprised of 1, 2 and n nucleosides, and the density was calculated under different numbers of bases codon. In the case of MS data storage, the capacity means the total number of peaks formed by the codons. Therefore, densities of various codons are received by that the accumulated number in Fig. 3a was converted into binary number and divided by the number of bases forming a codon. The information density under specific numbers of nucleosides with codons comprised of 1–4 bases is presented in Fig. 3b. A distinct increase in the information density is observed as the number of additional unnatural nucleosides increased from 0 to 60. However, a noticeable decline appears in information density when the number of bases that makes up a codon increases, except for no difference found between 1-base codon and 2-base codon. Theoretically, a 1-base codon with a library consisting of 60 unnatural nucleosides and 4 natural nucleosides offers an information density as high as 6 bits/nt. Furthermore, as more unnatural nucleosides join in, the density will increase. Although this mass-based coding density may not have a direct impact on the physical density of data storage, it should serve as a useful reference for understanding "molecular-weight" based DNA storage. For example, if MS technology progresses to higher sensitivity, deeper detection, and smaller physical volumes, as described in Fig. S2, it will become a key characteristic for "molecular-weight" based DNA data storage.

The capacity is another parameter essential for the evaluation of molecular-weight based MS DNA data storage, which means the maximum volume of data stored by a smallest unit of DNA. In other word, capacity stands for the number of the distinguishable peaks within a mass spectrum under various codon lengths and nucleosides types. Therefore, it is equal to the count calculated in Fig. 3a. The capacity of mass spectra in the presented data storage scheme is shown in Fig. 3c and Fig. S3. Capacities versus nucleosides were simulated with codons consisting of 2–7 bases, where the number of nucleosides is composed of 4 natural and 1 to 4 unnatural nucleosides. The information capacity rises with the inclusion of types of nucleosides for each base number of a codon, and the potential components with molecular weight differences

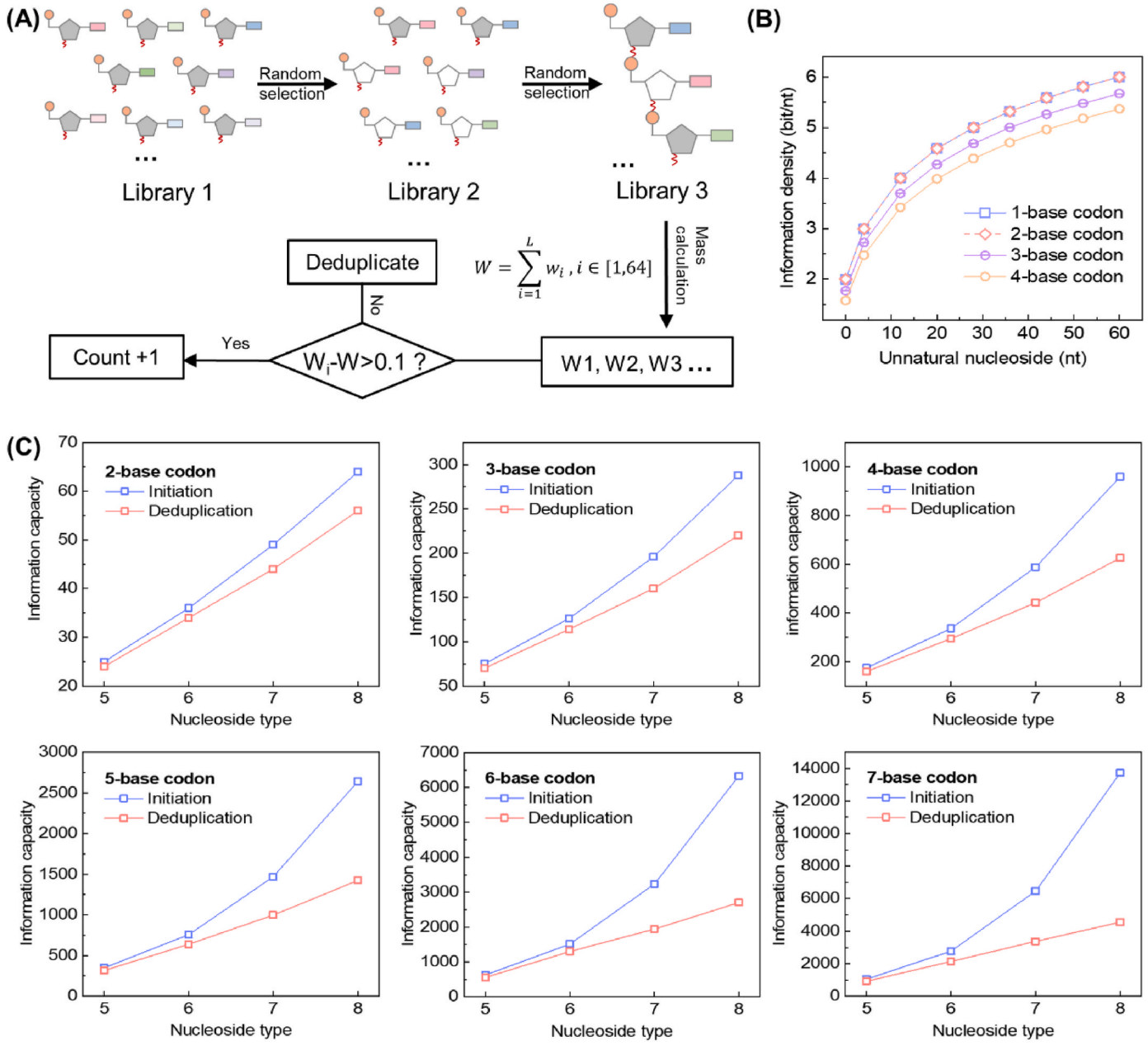

Fig. 3. Characteristics of MS-based DNA data storage. (a) Flowchart of DNA molecular weight-based data storage implemented by algorithm. Simulation results of (b) high-density information storage based on molecular weight and (c) high capacity of mass spectra for data storage.

smaller than 0.1 *m/z* are deduplicated (Fig. 3c). Correspondingly, when the type of nucleosides is fixed, the information capacity increases exponentially with the growing number of bases in a codon (Figs. S3a–d). Smoother increases are observed for deduplicated capacities compared to initial capacities. Despite the removal of molecules with identical mass values in mass spectrometry (MS)-based DNA data storage, the substantial increase in molecular diversity achieved through the incorporation of unnatural nucleosides provides a vast reservoir for data encoding. Furthermore, even though the growth in density and capacity is very slow after adding 60 unnatural nucleosides as discussed in Fig. 3, it can be further expanded slightly if the library 2 is enlarged by adding more unnatural nucleosides for MS-based data storage. However, it should be noted that the density and capacity reach a plateau unless the types of nucleosides grow exponentially as the density and capacity.

3.3. Medical data storage and retrieval by "multi-layer" data storage architecture

To verify the effectivity of the MS-based "multi-layer" encryption strategy, we stored Hippocratic oath by traditional DNA data storage platform and a medical examination report by the proposed method in this work. When encoding Hippocratic oath in 8 strands of 1400 nt DNA sequences inserted in plasmids using traditional DNA data storage scheme, the information is directly decoded using Sanger sequencing (Fig. S4). Considering the confidentiality of information, a medical examination report is stored and retrieved by the developed "multi-layer" encryption assembled DNA-based information storage. The achievement of "molecular weight"-based DNA data storage allows us to develop a "multi-layer" DNA encryption scheme. As explained above and illustrated in Fig. 1, a "three-layered" system was constructed. The medical examination report was first encrypted behind an IT layer to protect

patients' privacy, with a "key" generated using Blowfish. The secured information was then encoded in DNA sequences and synthesized as plasmids, which served as the second encryption layer. Meanwhile, the key was encoded using the molecular weights of oligonucleotides or nucleosides for top layer encryption. Specifically, the report is encrypted by Blowfish algorithm to generate a "key" and encrypted report. The encrypted report is further encoded into plasmid DNA with ensuring high storage stability and sequencing accuracy. Correspondingly, the "key" is set as "BELIEF" and stored using molecular weights of either oligonucleotides or nucleosides, which is read by MS. In this case, the report cannot be retrieved solely through sequences recovered by Sanger sequencing without the password for decryption.

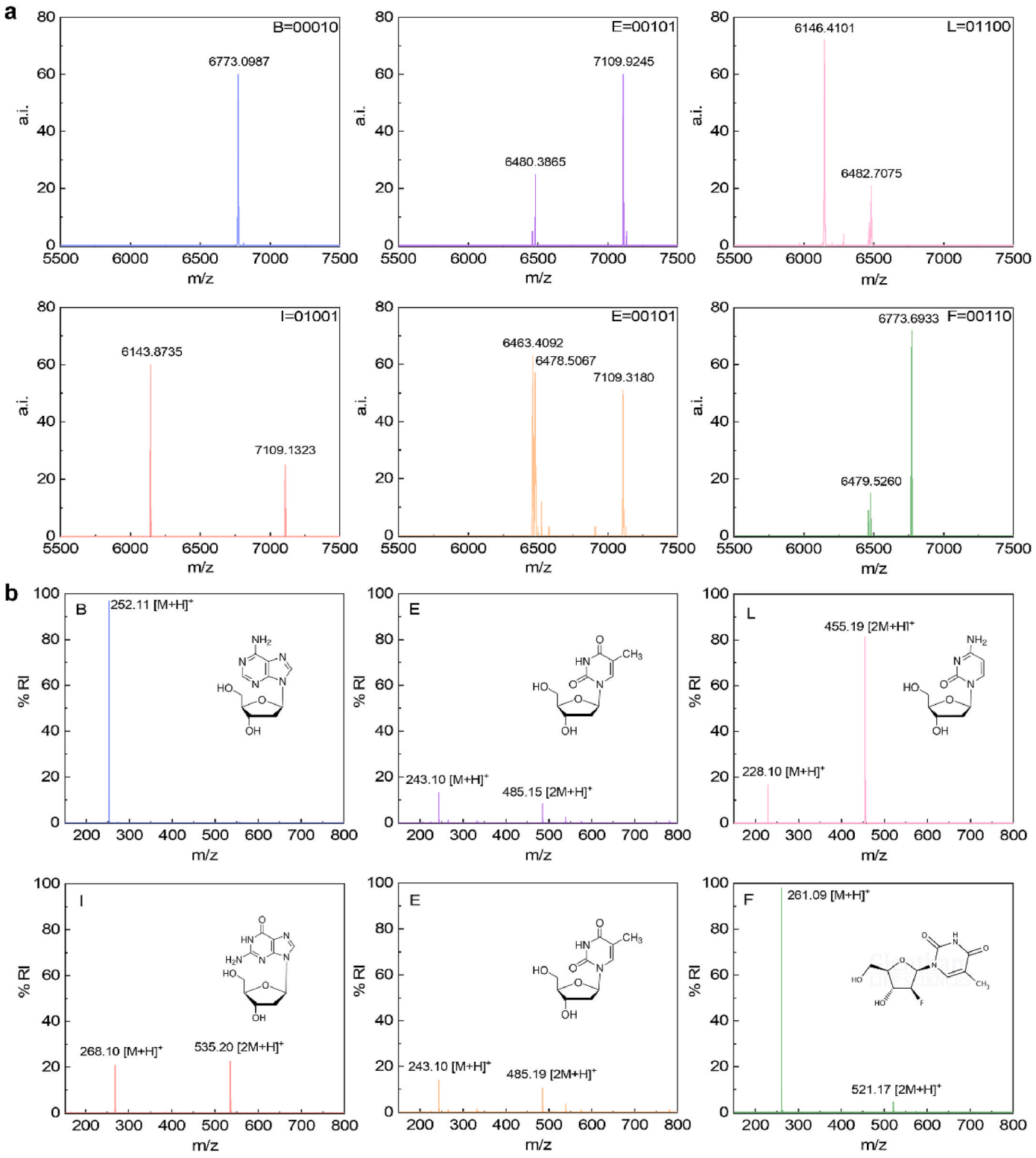

Fig. 4. Mass spectra for the "key" information retrieval. (a) Positive ion mode MALDI spectra for information stored by molecular weights of oligonucleotides. (b) Positive ion mode LC-MS spectra for information stored by molecular weights of nucleosides.

The resulting mass spectra for six components of the "key" are presented in Fig. 4a. Mass values in each spectrum could be traced back to each oligonucleotide, demonstrating a specific binary code. These binary codes were subsequently converted back into decimal digits and then mapped back to English letters, thereby reconstructing the complete "key". Alternatively, LC-MS was recorded in the case that "BELIEF" was stored by nucleosides, due to the inaccuracy of MALDI for objects with small molecular weights. Each of the six molecular weight was analyzed to extract the "key". The spectra corresponding to each sample at retention time of 2.63 min, 4.52 min, 0.91 min, 0.91 min, 4.55 min and 4.81 min are displayed in Fig. 4b. In these spectra, each molecular weight was associated with one of the nucleosides listed in the table of 26 nucleosides (Table S4). "BELIEF" was subsequently reconstructed by mapping each nucleoside back to its corresponding English letter.

Following a reverse process of Blowfish encryption, a fully recovery of medical examination report can only be achieved by integration of the "key" accessed by MS with the encrypted report, of which the encrypted report was decoded from 10 strands of 1500 nt DNA sequences inserted in plasmids by Sanger sequencing. Full coverage to each designed sequence was found for each information-encoded sequence, which was confirmed by alignment (Fig. S5). Furthermore, agarose gel electrophoresis was performed for information-encoded sequences to examine the possibility of data interpretation under low concentration following gradient dilutions of plasmid DNA template. PCR was conducted for each dilution, with the target products outlined in red. As depicted in Fig. 5a, target products were detected from Lane (1) to Lane (5), indicating that data retrieval is possible with an extremely low input amount of 10^{-9} ng. The gel image was normalized into countable intensity by ImageJ, and the relative intensities are presented in Fig. 5b. No significant drop in intensities at high input amount exceeding 10^{-3} ng, whereas an exponential decline was observed from 10^{-3} to 10^{-9} ng, and no amplified product was detected with a template amount of 10^{-12} ng. Additionally, Sanger sequencing was conducted for PCR product of 10^{-3} and 10^{-9} ng plasmid DNA, with a portion of peaks depicted in Fig. 5c. The copy number calculation is detailed in Supplementary note 2. Therefore, data access is achievable with single-digit numbers of copies of plasmid DNA. Fig. S6 illustrates the alignment of PCR products obtained from 10^{-3} ng and 10^{-9} ng of DNA to the inserted sequences in the plasmid. The alignment demonstrates 100 % full coverage, indicating the potential readability of the information, which is consistent with the agarose gel electrophoresis results.

3.4. Stability of storage media in the developed "multi-layer" DNA data storage

The stability of storage media was assessed through an accelerated aging procedure, followed by qPCR analysis of the encrypted information-encoded plasmid and ^{1}H NMR scanning for the "key" evaluation. Initially, qPCR was performed using 10-fold serial dilutions of template plasmid ranging from 10° to 10^{-7} ng, and the standard curve of Ct values versus DNA amount is presented in Fig. 6a. A linear equation of $y = 2.55307x + 21.41442$ ($R^2 = 0.939$) was derived and utilized to quantify the survival of the template plasmid after being exposed specific accelerated aging procedure. Agarose gel electrophoresis was then conducted for PCR product of plasmid samples before and after accelerated aging, with the results displayed in Fig. 6b. The plasmid post-

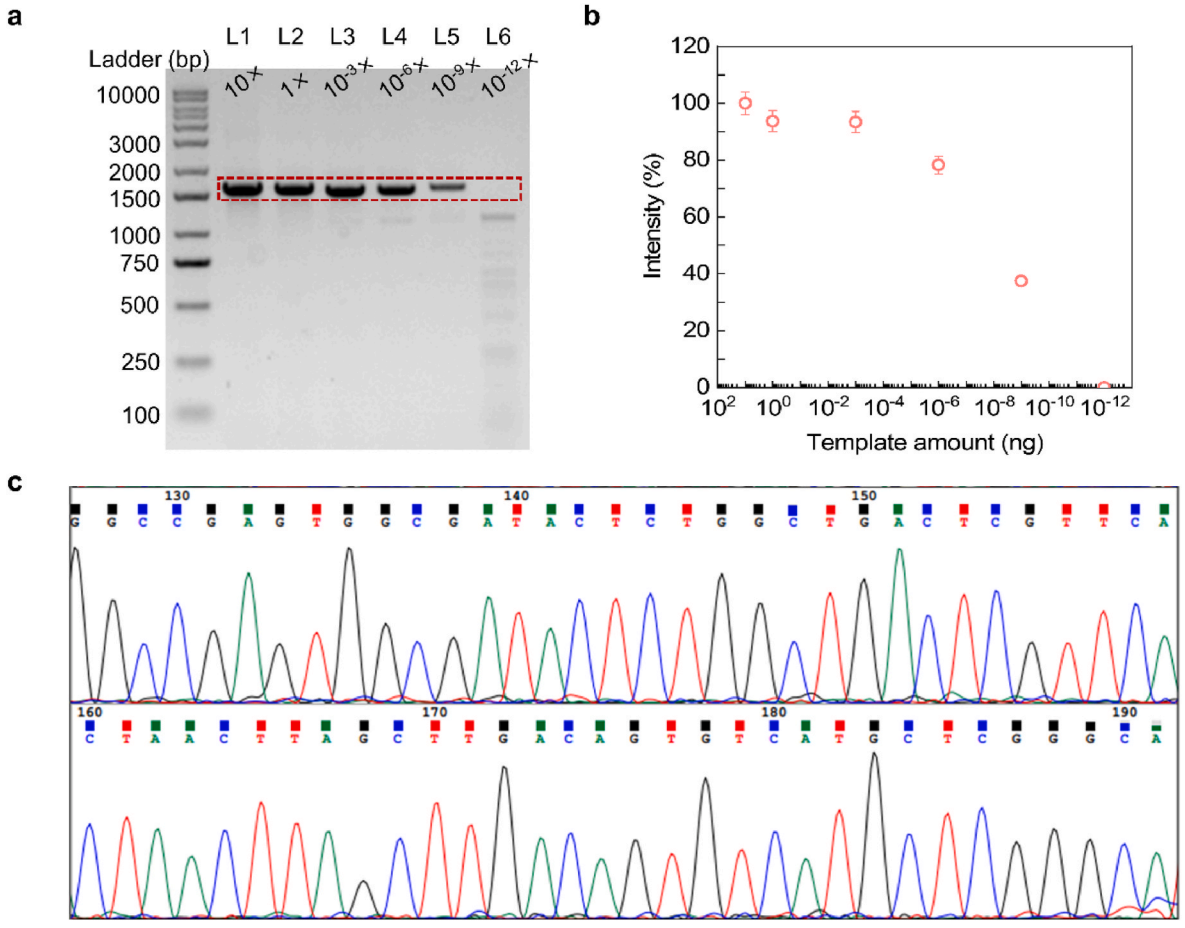

Fig. 5. Feasibility of the encrypted information retrieval. (a) Agarose gel image for PCR product using serial dilutions of template plasmid DNA. (b) Normalized readable trend by PCR amplification of plasmid DNA template with serial dilution. 1 × equals to 1 μL of 4 ng/μL plasmid. Intensity of 10 × sample is set as 1. Error bars: SD. n = 3. (c) Sanger sequencing from PCR product of 10^{-9} × information-encoded plasmid.

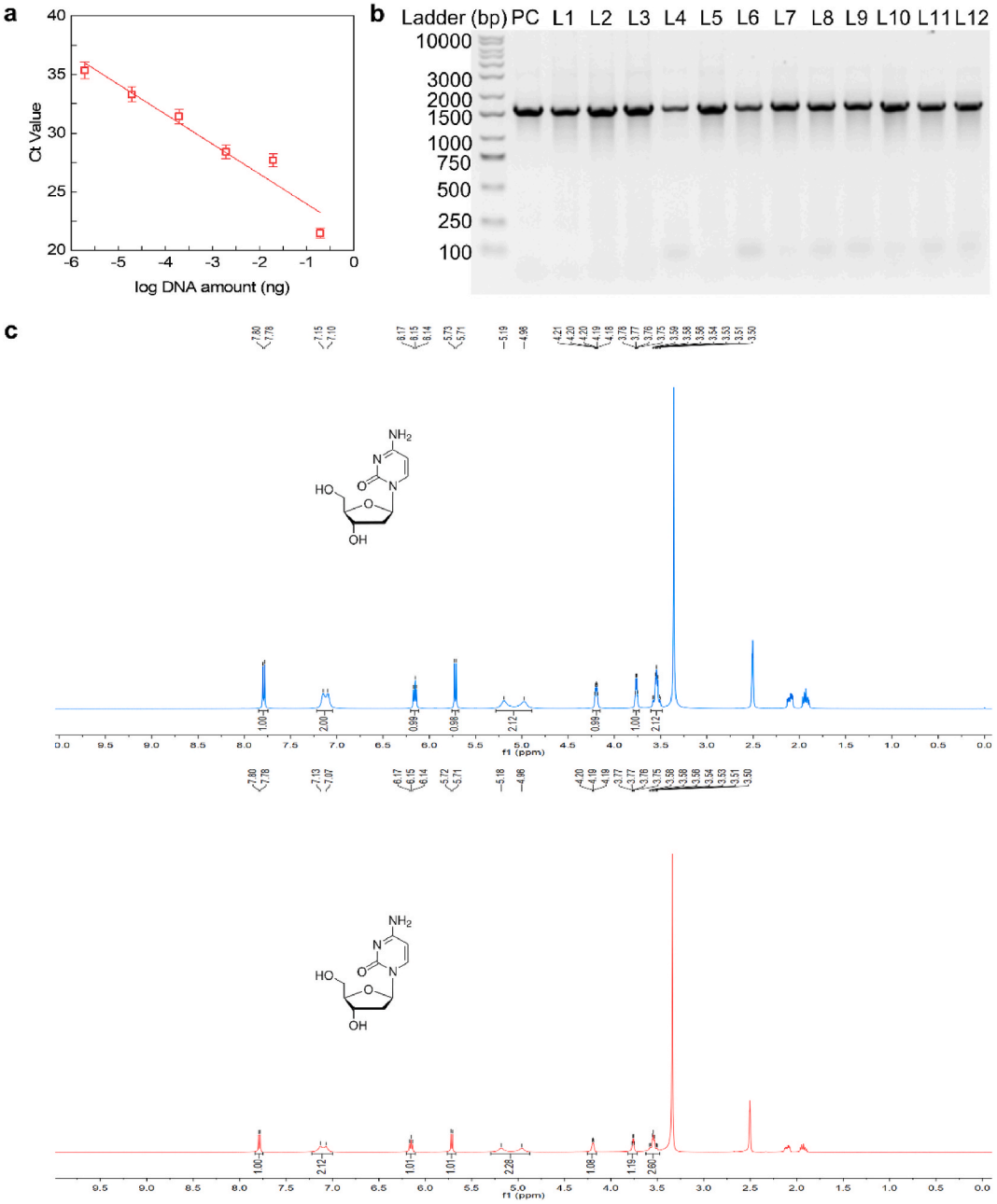

Fig. 6. Stability of BT layers in DNA data storage scheme. (a) Standard curve from plasmid template with series dilution from 10° to 10^{-7} ng. (b) Agarose gel image for PCR product of plasmids after aging at different temperatures after different days. PC: PCR product of plasmid before aging. L1-L12: PCR products of plasmids after aging at 60 °C (odd lanes) and 65 °C (even lanes) for 1, 3, 5, 7, 9, 11 days. (c) ^{1}H NMR spectra of dC before and after accelerated aging at 70 °C for 7 days. Error bars: SD. n = 3.

aging exhibited comparable length to those before the accelerated aging test, indicating a high information stability of the second BT layer. The unevenness of the intensity could be related to either accidental missing of plasmid during the accelerated aging procedure or evaporation during PCR. Besides, the plasmid after aging was subjected to qPCR and the resulting amplification curves are depicted in Fig. S7. A detailed explanation and calculation of the lifespan are provided in Supplementary note 3. The lifespan of the information-encoded plasmid was estimated to be 351.6 years, which is far more stable than the conventional storage devices, such as 10–20 years for magnetic tapes, 3–5 years for flash storage, 5–10 years for CD and DVD, etc. [26,27].

Given that the oligonucleotides degrade under elevated temperature and humidity, nucleosides are preferred for long-term information storage, whereas oligonucleotides are employed for temporary storage. The ^{1}H NMR spectra of nucleosides before and after accelerated aging are displayed in Fig. 6c and Fig. S8-13. No degradation is measured for natural nucleosides after aging at 70 °C for 7 days, let alone the more stable unnatural nucleosides. Besides, the plasmid was mixed with approximately 100 times the amount of nucleoside (one of the nucleosides encoding the key) and subjected to an accelerated aging procedure again, followed by mass spectrometry scanning without purification. As displayed in Fig. S14, the molecular weight of nucleoside was detected successfully after accelerated aging, indicating that plasmid degradation has not affected in the retrieval of data stored based on molecular weight under the tested conditions, while the amount of nucleoside is significantly greater than that of plasmid DNA. Therefore, storage media in mass and sequence content layers offer high stability, demonstrating sufficient performance of the "multi-layer" DNA data storage platform.

4. Discussion and conclusion

Despite that DNA data storage has been studied for decades, its concept has been largely restricted to storing information in DNA sequences alone. This limited the potential uses of DNA data storage and its further development. In this work, we extended the DNA data storage method and effectively established a "multi-layer" DNA data storage working scheme for medical data encryption by exploring the potential of DNAs' "molecular weight" for information storage. Medical examination report was firstly encrypted within an IT layer to safeguard the privacy of patients, with a "key" generated using Blowfish. Subsequently, the protected information was encoded in DNA sequences and embedded into plasmids because of their suitability for enhancing information safety, attributed to their storage stability and sequencing accuracy, which was considered as the second encryption layer. Meanwhile, the key was encoded based on the molecular weights of either oligonucleotides or nucleosides for top layer encryption. By these, our approach provides a useful way to protect any types of medical data as from information leakage or damage. The secure, long-term storage of medical data is critical, as it serves as the foundation for many key applications. Patients rely on the integrity and confidentiality of their health records to obtain personalized therapy, which ensures treatment plans are matched to their unique genetic traits and medical history. For researchers and physicians, collecting and analyzing extensive genetic sickness data across time provides critical insights into disease progression, inheritance patterns, and the development of targeted therapies. Furthermore, the preservation of genetic illness data across generations is very important for determining genetic susceptibilities and implementing preventive strategies in familial and population contexts. Although this "proof of concept" study only shows the storage of limited size of data, the key concept of "multi-layered encryption" presented here should serve as a helpful guide for the future development of technology used for practical medical data encryption, thereby avoiding the problem—as is noted, medical data leakage can have serious consequences that damage patient-doctor trust and result in financial losses. The proposed method offers an option and presents potential for safely archive medical data for over hundred years due to the "multi-layer" encryption and long lifespan of DNA molecules. Except for storing medical report in DNA with enhanced security, stability, and accessing convenience, the proposed "multi-layer" encryption-based DNA information storage demonstrates huge potential for the recording of wide variety of information types when high security is pursued, such as tuberculosis reports, military strategic maps, pregnancy reports, passport, etc. (Fig. 7). Overall, the proposed strategy offers high security by a novel concept based on multi-layer encryption and decryption in DNA data storage.

The success of our established "multi-layer" encryption system owes to the use of MS analysis of DNA's molecular weight. MS has long been regarded as one of the most versatile analytical tools for DNA molecules due to the ability for providing structural information of tiny amount of molecules and the compatibility with various separation methods, including liquid chromatography (LC), gas chromatography (GC), capillary electrophoresis (CE), and metal-assisted laser desorption/ionization (MALDI) [28,29]. These have enabled users to detect 10^{-4} Da difference by scanning the mass spectra with 1 mL of 10 ppm or 1 mg sample in solid form in general [30]. In the field of data storage, MS-based approach is particularly suitable for storing read-once information, such as the entry of confidential information. Therefore, this work incorporates MS, encryption algorithm to DNA-based digital information storage, offering potential for "multi-layer" encryption for data archiving. As a proof-of-concept experiment, mass spectra were scanned to retrieve the "key" generated by Blowfish algorithm, thereby circumventing the challenges associated with unnatural nucleic acids sequencing, which are typically time and cost-intensive processes. In addition, utilization of unnatural nucleosides and nucleosides expanded the molecular alphabet, leading to expanded capacity and a potential information density of 6 bit/nt (if calculated by sequence-based DNA data storage method) in molecular weight-based MS DNA data storage by simulation. Furthermore, introducing unnatural bases and sequences by molecular weight-based data storage boost encryption diversity in DNA data storage. Nevertheless, it should be pointed out that the current expense associated with DNA data storage is not cheap, primarily due to the high costs incurred in DNA synthesis. Nonetheless, advancements in enzymatic DNA synthesis and improvements in instrumentation are expected to reduce these costs in the foreseeable future [10]. For example, TdT enzyme-based DNA synthesis is currently extensively studied and developed both in general laboratories and several advanced enterprises. It is expected that the synthesis of one base pair will be as easy as sequencing of that in the near future and the total cost of DNA data storage will gradually approach the current storage with silicon-based medium, leading to an expanded usage of DNA as data storage medium. In addition, although MS holds considerable advantages for storing the essential key information, the current MS analysis needs expensive machine and its sample preparation might be laborious. However, it is believed that the further development of novel MS equipment to optimize its cost and sample preparation or additional

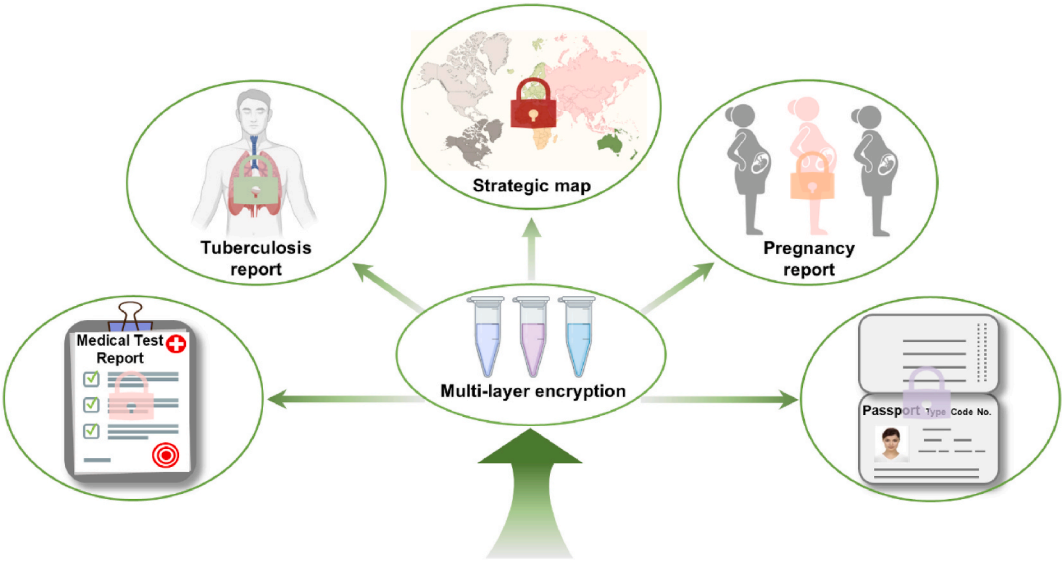

Fig. 7. Potential application scenarios of "multi-layer" encryption facilitated DNA data storage.

development of other "molecular weight" determination method should enable it to be used in more scenarios (Fig. S2).

The analysis of molecular weight by MS also enables nucleotides to be used for data storage, which should be particularly suitable for long-term secure data storage because they do not suffer from fragmentation or degradation with time like DNA long sequences or oligonucleotides [8]. Nucleosides can be used either directly or synthesized into DNA chains for data storage, attributed to the large library of molecules made up of natural and unnatural ones. Unlike the only 4 types natural nucleosides used for DNA data storage, there has already been nearly hundred types of unnatural nucleosides reported in RNA in the early 1990s [31]. In the current days, it is reported to have over 300 types of modification [32]. This large number of unnatural nucleoside types offers greatly expanded alphabet of molecules for encoding, thus providing improved information capacity. By the combination of DNA molecular weights, DNA sequence information and IT algorithm, we have built a "multi-layer" encryption way to securely store medical data. Losing any layer of information will lead to failure during decoding and decryption, which also enables it be better than other DNA encryption method to some extent (Table S6). This work thus opens a door for improving the data encryption by fully using the properties of DNA molecules. Inspired by the "multi-layer" encryption, more dimensional or layered encryption in DNA-based data storage can be applied in cases where security is paramount. This can be achieved by developing DNA-based data storage to higher flexibility using its chemical and physical properties at molecular level. Digital information can be encoded in the physicochemical properties of DNA molecules, such as DNA sequence length, conductivity, and UV or infrared absorbances, and then retrieved using fragment analyzers, electrical conductivity meters, and spectrometers. Aside from this, structural information, such as secondary structure and three-dimensional structure of DNA, may also be employed to capture crucial information. Furthermore, biological features such as regulated DNA sequences, protein recognition motifs, and enzymatic cutting sites of DNA may be employed to encrypt information. Moreover, other ways of encryption, e.g. L-DNA [33] or backbone of DNA [34] can be combined with our methods to further increase the encryption degree of the data.

To sum up, we have built a "multi-layer" encryption approach, which should provide a useful way for future highly-secure medical data archiving. By introducing "molecular weight" of DNA for data storage, we expanded the concept and potential of "DNA data storage" by introducing both natural and unnatural nucleosides for data storage. Our work also implies improving the capacity, efficiency, and diversity of DNA-based data storage architectures by introducing more possible DNA properties and analytical methods, showcasing significant potential for the development of convenient data retrieval tools such as ^{1}H NMR, fluorometer, fragment analyzer, and similar technologies.

CRediT authorship contribution statement

Jiaxin Xu: Investigation, methodology, validation, data curation, visualization, writing – original draft, writing – review & editing. **Yu Wang:** Methodology, software, data curation. **Xue Chen:** validation. **Lingwei Wang:** Resources. **Haibo Zhou:** Resources. **Hui Mei:** Supervision, funding acquisition, resources. **Shanze Chen:** Supervision, funding acquisition, resources. **Xiaoluo Huang:** Supervision, funding acquisition, resources, Conceptualization, writing – review & editing.

Declaration of competing interest

The authors declare no competing interests.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by National Key Research and Development Program of China (2021YFF1201700); National Natural Science Foundation of China (32201207, 32100734); Shenzhen Medical Academy of Research and Translation (C2302001; B2302041); Guangdong Basic and Applied Basic Research Foundation (2214050008970); Shenzhen Science Technology and Innovative Commission (KCXFZ202002011008256); Shenzhen Science and Technology Program (RCYX20221008092950122).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mtbio.2024.101221.

References

[1] E. Chisom, A. Becker, D. Heider, G. Hattab, Design considerations for advancing data storage with synthetic DNA for long-term archiving, Mater. Today Bio. 15 (2022) 100306.
[2] D. Reinsel, J. Gantz, J. Rydning. The digitalization of the world-From edge to core, 2018, pp. 1–33.
[3] S. Coughlin, D. Roberts, K. O'Neill, P. Brooks, Looking to tomorrow's healthcare today: a participatory health perspective, Intern. Med. J. 48 (1) (2018) 92–96.
[4] PwC, Driving the Future of, Health (2019) 1–24.
[5] G.M. Church, Y. Gao, S. Kosuri, Next-generation digital information storage in DNA, Science 337 (6102) (2012), 1628-1628.
[6] C. Bancroft, T. Bowler, B. Bloom, C.T. Clelland, Long-term storage of information in DNA, Science 293 (5536) (2001) 1763–1765.
[7] M.E. Allentoft, M. Collins, D. Harker, J. Haile, C.L. Oskam, M.L. Hale, P.F. Campos, J.A. Samaniego, M.T.P. Gilbert, E. Willerslev, The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils, Proc. R. Soc. A B 279 (1748) (2012) 4724–4733.
[8] M.E. Allentoft, M. Collins, D. Harker, J. Haile, C.L. Oskam, M.L. Hale, P.F. Campos, J.A. Samaniego, M.T.P. Gilbert, E. Willerslev, The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils, Proc. Biol. Sci. 279 (1748) (2012) 4724–4733.
[9] V.V.Z.D. Rasic, Semiconductor synthetic biology roadmap, Tech Rep. (2018) 1–32.
[10] The Future of DNA Data Storage. Potomac Institute for Policy Studies, 2018.
[11] V. Zhirnov, R.M. Zadegan, G.S. Sandhu, G.M. Church, W.L. Hughes, Nucleic acid memory, Nat. Mater. 15 (4) (2016) 366–370.
[12] Y. Hao, Q. Li, C. Fan, F. Wang, Data storage based on DNA, Small Struct. 2 (2) (2021) 2000046.
[13] C. Mao, S. Wang, J. Li, Z. Feng, T. Zhang, R. Wang, C. Fan, X. Jiang, Metal–organic frameworks in microfluidics enable fast encapsulation/extraction of DNA for automated and integrated data storage, ACS Nano 17 (3) (2023) 2840–2850.
[14] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E.M. LeProust, B. Sipos, E. Birney, Towards practical, high-capacity, low-maintenance information storage in synthesized DNA, Nature 494 (7435) (2013) 77–80.
[15] R.N. Grass, R. Heckel, M. Puddu, D. Paunescu, W.J. Stark, Robust chemical preservation of digital information on DNA in silica with error-correcting codes, Angew. Chem., Int. Ed. 54 (8) (2015) 2552–2555.
[16] Y. Erlich, D. Zielinski, DNA Fountain enables a robust and efficient storage architecture, Science 355 (6328) (2017) 950–954.
[17] J. Koch, S. Gantenbein, K. Masania, W.J. Stark, Y. Erlich, R.N. Grass, A DNA-of-things storage architecture to create materials with embedded memory, Nat. Biotechnol. 38 (1) (2020) 39–43.
[18] Y. Ren, Y. Zhang, Y. Liu, Q. Wu, J. Su, F. Wang, D. Chen, C. Fan, K. Liu, H. Zhang, DNA-based concatenated encoding system for high-reliability and high-density data storage, Small Methods 6 (4) (2022) e2101335.
[19] Y. Zhang, L. Kong, F. Wang, B. Li, C. Ma, D. Chen, K. Liu, C. Fan, H. Zhang, Information stored in nanoscale: encoding data in a single DNA strand with Base64, Nano Today 33 (2020) 100871.
[20] K. Assa-Agyei, F. Olajide, A comparative study of twofish, blowfish, and advanced encryption standard for secured data transmission, Int. J. Adv. Comput. Sci. Appl. 14 (3) (2023) 393–398.
[21] M.N. Khatri–Valmik, V.K. Kshirsagar, Blowfish algorithm, IOSR J. Comput. Eng. 16 (2) (2014) 80–83.
[22] A. Alabaichi, Faudziah Ahmad, Ramlan Mahmod, Security analysis of blowfish algorithm, in: Second International Conference on Informatics & Applications (ICIA), IEEE, 2013, pp. 12–18.
[23] X. Huang, J. Cui, W. Qiang, J. Ye, Y. Wang, X. Xie, Y. Li, J. Dai, Storage-D: a user-friendly platform that enables practical and personalized DNA data storage, iMeta (2024) e168.
[24] M. Blawat, K. Gaedke, I. Hütter, X.-M. Chen, B. Turczyk, S. Inverso, B.W. Pruitt, G. M. Church, Forward error correction for DNA data storage, Procedia Comput. Sci. 80 (2016) 1011–1022.
[25] Z. Ping, S. Chen, G. Zhou, X. Huang, S.J. Zhu, H. Zhang, H.H. Lee, Z. Lan, J. Cui, T. Chen, W. Zhang, H. Yang, X. Xu, G.M. Church, Y. Shen, Towards practical and

robust DNA-based data archiving using the yin-yang codec system, Nat. Comput. Sci. 2 (4) (2022) 234–242.
[26] Available from: https://www.arcserve.com/blog/data-storage-lifespans-how-long-will-media-really-last, 2024.
[27] C. Xu, C. Zhao, B. Ma, H. Liu, Uncertainties in synthetic DNA-based data storage, Nucleic Acids Res. 49 (10) (2021) 5451–5469.
[28] E. Esmans, D. Broes, I. Hoes, F. Lemiere, K. Vanhoutte, Liquid chromatography–mass spectrometry in nucleoside, nucleotide and modified nucleotide characterization 794 (1–2) (1998) 109–127.
[29] J. Tost, I.G. Gut, DNA analysis by mass spectrometry-past, present and future, J. Mass Spectrom. 41 (8) (2006) 981–995.
[30] Y.H. Lai, Y.S. Wang, Advances in high-resolution mass spectrometry techniques for analysis of high mass-to-charge ions, Mass Spectrom. Rev. 42 (6) (2023) 2426–2445.
[31] P.F.C.Patrick A. Limbach, A. James, McCloskey, Summary: the modified nucleosides of RNA, Nucleic Acids Res. 22 (12) (1994) 2183–2196.
[32] Available from:: https://www.bocsci.com/nucleosides-list-2000.html.
[33] C. Fan, Q. Deng, T.F. Zhu, Bioorthogonal information storage in L-DNA with a high-fidelity mirror-image Pfu DNA polymerase, Nat. Biotechnol. 39 (12) (2021) 1548–1555.
[34] C. Pan, S.K. Tabatabaei, S.M.H. Tabatabaei Yazdi, A.G. Hernandez, C.M. Schroeder, O. Milenkovic, Rewritable two-dimensional DNA-based data storage with machine learning reconstruction, Nat. Commun. 13 (1) (2022) 2984.