creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학박사 학위논문

High Information Capacity and Low Cost DNA-based Data Storage through Additional Encoding Characters

인코딩 문자 추가를 통한 고효율 저가격 DNA 기반 정보 저장법에 관한 연구

2018 년 8 월

서울대학교 대학원

전기컴퓨터 공학부

최영재

High Information Capacity and Low Cost DNA-based Data Storage through Additional Encoding Characters

지도 교수 권 성 훈

이 논문을 공학박사 학위논문으로 제출함
2018 년 8 월

서울대학교 대학원
전기컴퓨터 공학부
최 영 재

최영재의 공학박사 학위논문을 인준함
2018 년 8 월

위 원 장 이 신 두 (인)

부위원장 권 성 훈 (인)

위 원 윤 성 로 (인)

위 원 방 두 희 (인)

위 원 박 욱 (인)

Abstract

Storing digital data in DNA is the process of encoding digital data into DNA sequences, synthesizing and storing these. Recently, the platform has been emerged with the possibility to supplement the current backup data storage with infrequent access, due to its physical advantages compared to conventional storage media. First, DNA can be maintained for centuries, which is in contrast to conventional storage media that require power supply or be rewritten for data retention. Second, DNA has physical information density that can store hundreds of petabytes (PB, 10^{15} bytes) per gram, thousands of times higher than conventional storage method. The major goal of previous research on DNA-based data storage was to improve data encoding algorithms for reducing data error or loss. Design rules for Data to DNA encoding and error correction functions were suggested.

The next step towards DNA-based data storage is to reduce the cost for storing the data and enable the practical use. Current cost for DNA-based data storage is about 3500 USD per storing 1 MB of data storage. As a first step to practical implementation, this dissertation shows the possibility of reducing the cost of DNA- based data storage by 50% by increasing the amount of data that can be stored per synthesized DNA, i.e., the information capacity, above the previous theoretical maximum. The proposed idea is to use degenerate bases,

which are mixes of the four encoding nucleotides, as additional encoding characters with the DNA encoding characters A, C, G and T. I propose a completely novel approach utilizing a synthetic process, whereas the existing studies were algorithmic optimizations and simple demonstrations.

Using the proposed idea, I demonstrated and simulated the total process of the DNA-based data storage, including Data to DNA encoding, molecular biology-based DNA handling and DNA sequence to Data decoding. From this, the theoretical maximum information capacity, which is equivalent to $\log_2$ value of the number of encoding characters, is increased from $\log_2 4$ to $\log_2 15$ (bit/nt) by adding 11 degenerate bases to the original four encoding characters. The DNA length required for storing data was experimentally reduced by more than half compared to that of the 4 character-based system. Also, from the simulation and cost projection, the cost of storing 1 MB is projected to be reduced by 50% compared to the previous cost. The data writing or DNA synthesis cost is decreased because the length of DNA required to store data is reduced to less than half.

Since the method only needs minor modifications of the encoding and DNA synthesizing processes, it can be applied to nearly all proposed DNA-based data storage methodologies and could increase the economic efficiency. Therefore, it is expected that the proposed idea and the demonstration could be utilized for practical implementation of DNA-based data storage.

Keyword: DNA-based data storage, Data Storage, Degenerate Base, Molecular biology, Synthetic biology

Student Number: 2012-23246

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

In this chapter, increasing demand for data storage and the DNA-based data storage that currently being studied as an alternative storage method will be described. After that, information capacity, which is essential for practical use of the DNA-based data storage will be introduced. Finally, the subject of this dissertation, new methodology of DNA-based data storage with increased information capacity will be presented.

1.1. Increasing Demand for Data Storage

The research group IDC predicted that the human-generated data is growing exponentially every year, reaching 163 zettabytes (ZB, 10^{21} bytes) in 2025, which is ten times the 16.1ZB of data generated in 2016 (Figure 1.1 (a))[1]. Also, about 30% of the data is related to the continuity of our daily life, and it is critical to areas such as the medical application and commercial air travel. They must be stored on a variety of media and, if necessary, additional backups are also required. However, because the amount of data produced is so large and rapidly growing, there are technical limitations in conventional storage media developed to date. For example, in 2040, researchers forecasted that memory demand exceeds the global silicon supply, assuming that all memory is stored in flash-based memory for instant access[2].

About 50% of this surplus data is classified as Cold data[1], [3]. Cold data is infrequently accessed data that includes the historical data, or backup of the photo or document that people have used. An extreme example of this cold data is 'The Square Kilometer Array (SKA)' of the NASA. According to NASA's article (https://www.skatelescope.org/amazingfacts/), this multi-radio telescope built in Australia and South Africa will generate 10 times as much data as the global internet traffic every day. In case of cold data, different criteria are required for storage when compared to the frequently accessed data. As shown in Figure 1.1 (b), the expected storage life for cold data should be longer

compared to that of other types of data. Also, there should be minimum maintenance for cost saving such as electricity and space. Also, since the data is not accessed frequently, the access speed does not need to be fast[4].

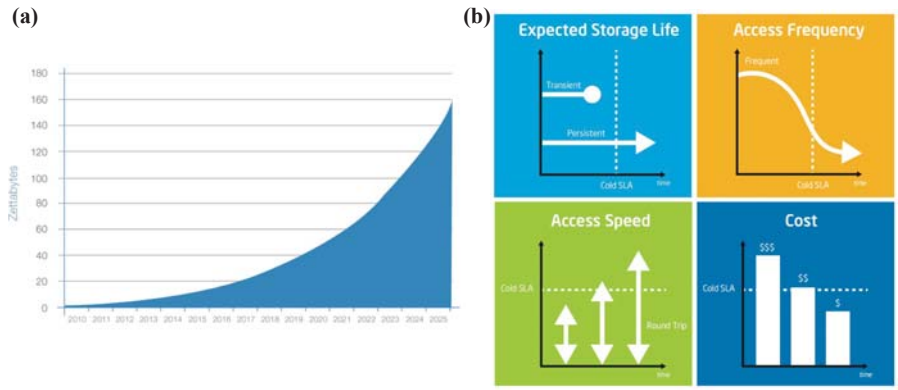


Figure 1.1 Increasing demand for data storage: (a) Annual size of the data created. The values up to the year of 2016 are actual values, and the values after that are predicted by the IDC[1]. Zettabyte: 10^{21} bytes; (b) The requirements model for cold data storage. SLA means service level agreement between the customer and customer and service provider[4]. The figure has been modified from the previous research[1], [4].

1.2. DNA-based Data Storage

1.2.1. DNA as the Nature's Data Storage Medium

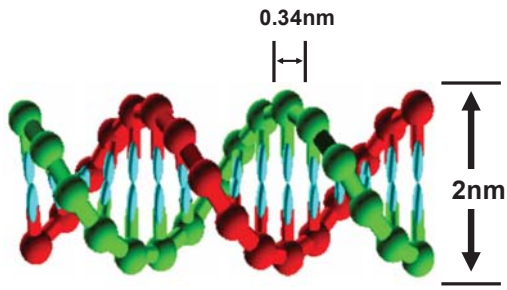


Figure 1.2 Structure and size of double-stranded DNA.

Before mankind store information, nature's life saved information, copied it, shared it with cells, and transmitted it to posterity. Nature stores genetic information using nucleic acid, usually DNA. DNA consists of two single strands hybridized to form a double strand, and the DNA consists of a 2 nm diameter, 0.34 nm high cylindrical base connected to each other (Figure 1.2). In humans, the length of the genome is about 3 billion base pairs, and if one base can encode 2 bits, the genome can hold data equivalent to a text file of 725 MB. Human cells store this data in nuclei with a diameter of about 6μm. It is capable of storing 5 petabits per square inch and is a thousand times the storage density of hard drives that are currently in commercial use. In addition, the DNA replication and error recovery is performed by enzymes with a size of several tens of nanometers or less. In this process, negligible amount of energy

is used, when compared to that from other systems. From this, DNA is the most energy-efficient data storage platform with highest data density.

1.2.2. DNA-based Data Storage

DNA-based data storage is a new concept of the data storage that converting digital data of 0,1 into a DNA base of A, C, G, and T, synthesizing the DNA sequence and used as a storage medium (Figure 1.3(a),(b))[2], [5]–[11]. The structure of a typical DNA-based data storage is shown in Figure 1.2 (b). Because of the limited length that can be synthesized in DNA synthesis technologies and sequenced in DNA sequencing technologies, information cannot be stored in a single DNA molecule. Instead, the information is divided into several pieces and an address is assigned. Both ends of the DNA are attached with sequencing and adapters for DNA amplification.

DNA-based data storage has two major advantages when compared to existing data storage media. First, the retention time of DNA is very long, and no other treatment for maintenance is required during storage. For example, 43,000 years old DNA from remains of a woolly mammoth extracted in high quality from ice[12]. Aside from extremely low-temperature condition like ice, DNA is reported to be intact for more than 2000 years in 18 °C with chemical treatment[13]. As can be seen in Figure 1.3 (b), the duration is orders of magnitude longer than that of existing data storage media[14]. In addition, if the DNA is properly sealed in the container, there is no need for additional

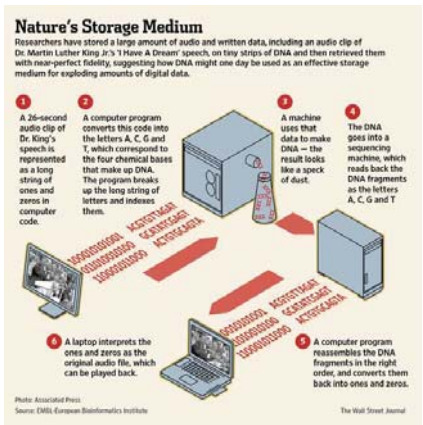
maintenance or electricity supply for storage, which is advantageous over existing media that require electrical supply. Second, DNA has high physical information density of petabyte (PB, 10^{15} bytes) per gram. Ideal DNA-based data storage is to store binary data (0,1) in a single DNA base position (A, C, G, or T) and single base pair with a volume of 0.3 nm tetrahedron can store 2 bits. This means that DNA has about thousand times higher information density compared to flash memory and about one million times higher physical information density than a hard disk (Figure 1.3 (b)). Based on these advantages, DNA-based data storage is expected to be used in cold data storage in near future.

The concept of DNA-based data storage has been proposed since early 2000[15]. Even when the synthesis and analysis methods of DNA used for storage were not established, the structure of the storage method and its advantages were first introduced. After that, actual implementation of the storage method has been demonstrated in the year of 2012 as the throughput and price of DNA synthesis and analysis have recently dropped[5].

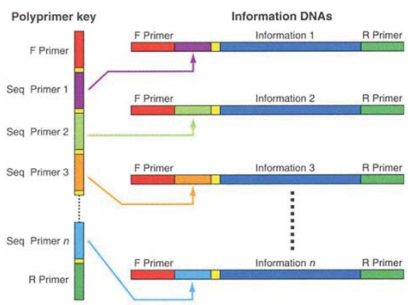
However, DNA-based data storage has cost problem before it can be practically used. This is due to the high cost of reading and writing data when using DNA. The process of writing data to DNA includes the encoding of DNA and the chemical synthesis of this DNA, and the reading process includes DNA reading using next-generation sequencing (NGS) and its computational decoding. Among these, the major percentage of the cost is the data writing.

Currently, the storage cost of storing a 1MB using DNA-based data storage is about 3500 USD and 99%, of which is the cost of data writing. Previous studies have suggested that DNA-based data storage can be used practically when data writing prices are about 100 times cheaper[6]. A more detailed analysis of the amounts of the cost will be described in Chapter 5.

(a)



(b)



(c)

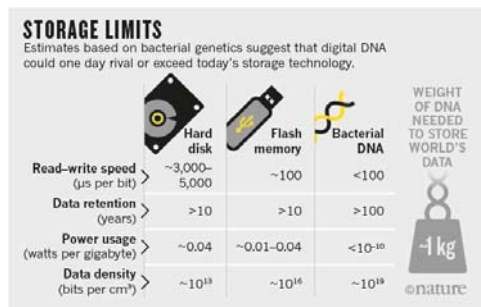


Figure 1.3 Concept and advantage of DNA-based data storage: (a) Concept of DNA-based data storage[16]; (b) Comparison of DNA -based data storage with other storage media in terms of read-write speed, data retention, power usage and physical information density (data density in figure)[14]. The figure has been modified from the previous research articles [14], [16].

1.2.3. Information Capacity of DNA-based Data Storage

Currently, approaches to increase the practicality of DNA-based data storage include decreasing the cost of DNA synthesis and increasing information capacity. In this dissertation, I focused on the increase of information capacity among these two methods. Information capacity is the amount of information that can be stored in one base position. Thus, if the information capacity increases, the length of the DNA to be synthesized when storing the data is reduced, thereby reducing the price.

However, DNA-based data storage, which converts digital data of binary form 0,1 into DNA 4-base system consisting of A, C, G, and T, has a limit of information capacity of 2bit / nt. The information capacity that can be achieved if encoding with system of N encoding characters (or base) consisting of binary system is as follows.

$$Information\ capacity = log_2 N\ \ (bit/nt)$$

In addition, as described in Chapter 2.1, the highest information capacity of the previous studies achieved 1.57 bit /nt, so dramatic cost reduction is not possible even if information capacity of 2.00 bit/nt is achieved. In this dissertation, I propose a method to increase information capacity by adding additional encoding characters other than A, C, G, and T to data to DNA encoding.

1.3. Main Concept: Addition of Degenerate Bases to DNA-based Data Storage for Higher Information Capacity

As described in the last section, DNA-based data storage has advantages that traditional storage media does not have, but it is problematic because of its high cost. For practical implementation of DNA data storage, there is a need for research on DNA-based data storage that focuses on the fundamental problem of reducing the cost of all aspects of this technology. To meet this demand, by increasing information capacity, the goal of this thesis is to decrase the data writing price, which is the biggest portio of the cost of DNA-based data storage.

This dissertation aims to increase the information capacity of DNA-based data storage by using additional base for data to DNA encoding (Figure 1.4 (a)). The required condition for additional encoding character used is; 1) synthesis and analysis of encoding character should be possible with current technology and 2) cost of encoding characters should be same as current bases. In this dissertation, I use degenerate bases as a candidate. Degenerate base is defined as a mixture of A, C, G, and T at a specific base position. For example, if a particular DNA sequence is ‘CWA’ and W is a degenerate base with mix of A and T, then the DNA molecule of ’CAA’ and ‘CTA’ are presented. If 11 degenerate bases are added to four base types and used for encoding, the theoretical information capacity is increased by about 2 times to $\log_2 15 = 3.90$ bit/nt (Figure 1.4(b)). As a result, the length of the DNA to be synthesized

through the specific data is reduced to half, the DNA synthesis price is also reduced by half. In addition, it is possible to create different degenerates by adjusting the ratio between degenerate base ratios and it is possible to achieve infinite information capacity by separating them through analysis.

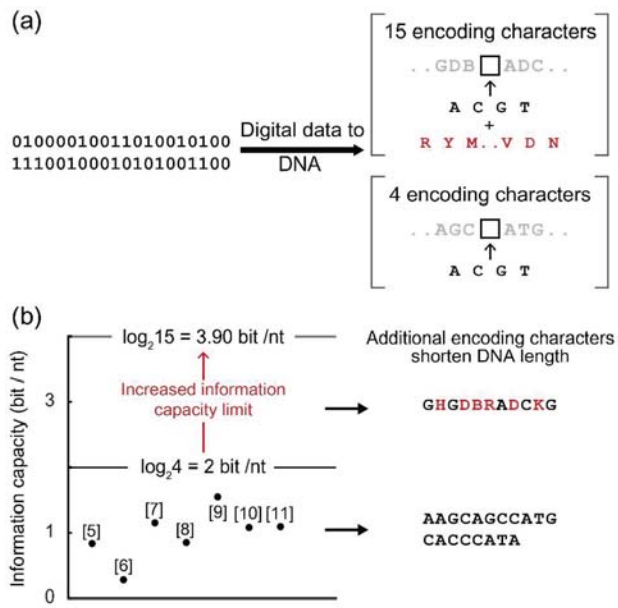


Figure 1.4 The main concept and the resulting increase in information capacity: (a) Adding the degenerate base(red) as extra encoding characters to digital data to DNA encoding; (b) Information capacity limit is increased from previous 2.0bit/nt to 3.90bit/nt and DNA length for storing the specific data ia shortened. The dots in the graph describe the information capacity in previous research, and the numbers indicate the corresponding reference. This figure has been modified from the previous research[17].

1.4. Outline of the Dissertation

In this dissertation, the concept is verified through both wet-lab experiment and simulation. In Chapter 2, previous DNA-based data storage methods are described and compared in terms of information capacity and cost. After that, the detailed introduction of degenerate bases is followed. In Chapter 3, a demonstration of the concept, from design of the storage and molecular biology-based experiment for handling and sequencing DNA to DNA to data decoding is covered. In Chapter 4, a simulation work to see if it could be actually scaled up will be demonstrated. Finally, Chapter 5 discusses how the proposed approach can actually improve the field by improving information capacity and cost, and how it can be developed in the future.

Chapter 2. Background of the Dissertation

In this chapter, an introduction of previous DNA-base data storage methods including data to DNA encoding algorithm, error correction code are covered. Also, comparison of these methods regarding information capacity and cost is described. Finally, the detailed introduction of degenerate bases as additional encoding characters is followed.

2.1. Previous DNA-based Data Storage Methods

2.1.1. The Nature of DNA to be Considered as Storage Media

In order to use DNA in data storage, it must reflect the constraints that arise in synthesizing and manipulating it. The constraints that occur in these two processes are the main source of errors in data storage and correcting or avoiding them should be reflected in the encoding algorithm of the DNA-based data storage.

Currently, the phosphoramidite method is generally used to synthesize DNA as oligonucleotide form. In this method, the designed DNA sequence is synthesized by serial chemical linking of the phosphoamidite corresponding to one nucleotide from 3 prime to 5 prime(Figure 2.1)[18]. From this, since the probability that each nucleotide can be successfully chemically linked is 99.5%, a yield of about 30% is obtained in synthesizing an oligonucleotide of about 200 nt. In addition, DNA that fails to link can be filtered by various methods of purification, but a large amount of DNA is left and a error occurs as the length of DNA is shorter than the design (deletion). In addition, unnecessary nucleotides are allocated in the linking process, or additional DNA is attached to cause insertion or substitution errors. Of these, the major error is deletion, which is about 50% of the total DNA molecule when synthesized at about 200 nt. In order to solve this problem, the synthesis length of the DNA should be reduced to about 150 or a new synthesis technique should be developed.

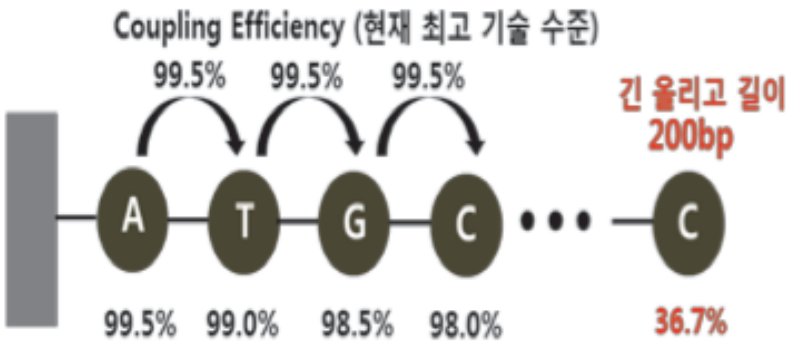


Figure 2.1 A simplified diagram of the DNA synthesis method.

Since the oligonucleotide is a single strand form when synthesized, it is necessary to carry out an amplification process using DNA polymerase (PCR, Polymerase Chain Reaction). This amplification is also necessary when copying data or when increasing the amount for analysis. During the amplification, the homopolymer (i.e. repetition of a specific DNA sequence such as AAAAA...) or GC contents (the ratio of base G and C in the entire sequence) in the DNA sequence in this process causes the amplification to be inefficient[19], [20]. This inefficient amplification will only occur in a small number of fragments in DNA library, resulting in un-even profile between fragments during the amplification process and leading to loss of data. In addition, in certain NGS methods homopolymer DNA exhibit a high error rate during sequencing[21]. Generally, it is known that amplification efficiency is guaranteed when the homopolymer length is 3 nt or less and the GC contents are between 30% and 60%. Previous studies have shown that very low

amplification efficiencies are found in other GC contents[19]. This is because small amounts of efficiency change can be seen to be very large in a tens of cycles of the PCR cycle since other molecules increase exponentially in the process.

In addition, if there is a high similarity between DNA molecules, interactions between molecules can lead to unwanted DNA assemble or low amplification efficiency. As a method to solve this problem, it has been proposed to amplify single molecules into wells or emulsions[22].

2.1.2. Data to DNA Encoding Algorithms

The first step in DNA-based data storage is to convert binary data consisting of 0,1 into DNA bases consisting of A, C, G, and T. The most basic encoding method is to convert 2-bit digital data to 4 variables, A, C, G, and T, such as 00=A, 01=C, 10=G, and 11=T, but this method is not actually used since the homopolymer could be generated and GC contents could not be controlled. To solve these problems, various data to DNA encoding algorithms have been proposed. First, an encoding method based on random DNA generation has been introduced. George Church group proposed an encoding method that randomly matches 0 to A or C and 1 to T or G[5]. From this method, even if there is data to be repeated, the homopolymer does not occur. Also, DNA fountain, which extracts a fragment of data according to the random seed, encode 00, 01, 10, and 11 corresponding to A, C, G, and T, respectively, and

discard it if homopolymer or high GC contents and re-extracting the fragment according to the seed, was proposed (Figure 2.2 (a))[10].

Also, rather than a method based on this random extraction, it was also suggested to create a DNA codon made of 3 bases and match the data (Figure 2.2 (b))[7]. This method places the last base on the DNA codon differently than the previous one so that no more than 3 bases of homopolymer are generated when the codons are connected. All encoding methods were designed to control the length of the homopolymer in the designed DNA fragment, but only the DNA fountain technique was able to control the GC content.

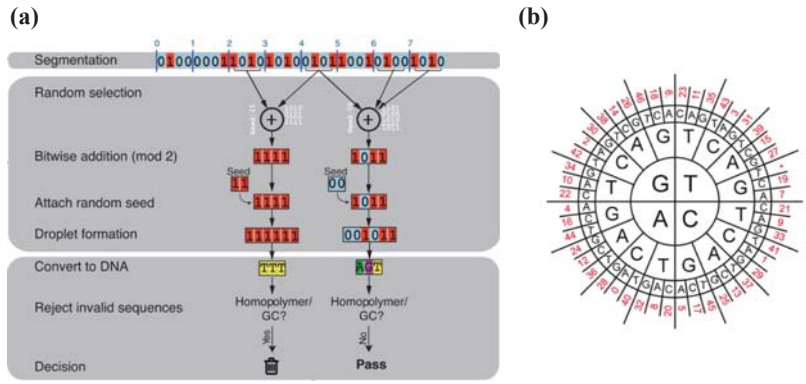


Figure 2.2 Various data to DNA encoding algorithms: (a) Schematic of the DNA fountain[10]; (b) DNA codons made with the DNA wheel and corresponding 47-digit numbers[7]. The figure has been modified from the previous research[7], [10].

2.1.3. Error Correcting Methods for DNA-based Data storage

DNA libraries are decoded through the process including synthesis, amplification, and sequencing, and the entire process has certain amounts of errors. In general, the error rate of Microarray-based DNA synthesis used in DNA-based data storage is 1% and error rate of in NGS is 0.1%[23]. Also, in the synthesis and amplification process, fragments may be lost in the library as mentioned in previous chapter.

The first approach to this error is to increase the copy of DNA per a fragment. This method can reduce the number of amplification, so even if there is uneven amplification efficiency between fragments, it can minimize fragment loss. In addition, even if there are errors that can occur during

synthesis and amplification, information of additional copies can be made to compare the same data to reduce errors. In the same vein, it is also possible to correct the error by making the amount of DNA read through sequencing much larger than the number of fragments.

By another approach, researchers attempted to correct errors by an algorithmic approach (Figure 2.3). The most basic way to do this is to create repetition by putting the same data into multiple fragments (Figure 2.3 (a))[6]. In this method, the ability of error correction varies depending on how many times the specific data is repeated. However, in this simple repetition, the same DNA sequence is present in multiple fragments, which can lead to unwanted hybridization between the fragments in the amplification process of the DNA. Also, if a particular DNA sequence pattern has low efficiency for amplification, the data correspond to the pattern can be lost, even in multiple fragments. To solve this problem, methods for creating new fragments through computation of different fragments have been introduced. As shown in Figure 2.3 (b), creating a new fragment by XOR calculation the data of two different fragments was proposed[8]. A Reed-Solomon error correction was also introduced, in which redundancy data is generated by multiplying data by a matrix (Figure 2.3 (c), (d))[7]. In the case of the XOR technique, two pieces of the three fragments are used to recover the data. However, the error correction capability cannot be adjusted. For the Reed-Solomon technique, the error correction capability can be adjusted according to the length of the redundancy such as:

$$2e + f \leq length\ of\ the\ redundancy$$

$$e: number\ of\ errors, f: number\ of\ erasures.$$

Depending on the ratio of the size of the Reed-Solomon block to the information block size, the error correction capability and the information capacity are traded-off. For the experiments reported so far, researchers have experimentally confirmed the optimal amount of error correction.

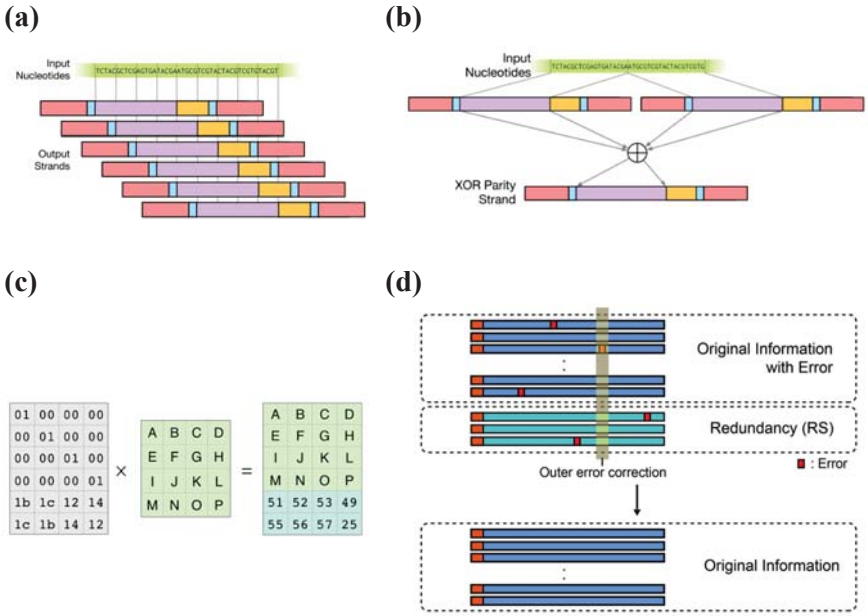


Figure 2.3 Various error correction for DNA-based data storage: (a) Simple redundancy-based error correction design[6]; (b)XOR based error correction design[8]; (c)Redundancy generation based a Reed-Solomon error correction method; (d) Error correction skimetic based a Reed-Solomon method[7]. The figure has been modified from the previous research[7], [8].

2.1.4. Comparison of DNA Storage Encoding Schemes and Experimental Results

Table 2.1 summarizes the major researches on DNA-based data storage proposed. As can be seen in the table, the early studies suggested had no error correction function, or even there was, it failed to recover the data completely. Since the stability of information storage has been established, various studies have been carried out to scale up DNA-based data storage to store large amounts of data or to increase its efficiency.

There are two criteria for the efficiency. The first is how much information (bit) can be put into a designed nucleotide(nt) or DNA base. This is called information capacity in this dissertation and is expressed in units of bit / nt. The total data stored is divided by the total number of nucleotides designed. The second criterion is how much data can be stored in unit weigh. This is called physical information density and is expressed in byte/g. In the ideal situation, only a single DNA molecule per fragment designed is needed for storage, so multiplying the information capacity by the molecular weight of the nucleotide could yield the value. However, due to the loss of DNA while amplification and the error rate of synthesis, data recovery is not possible when stored as monomolecules, and hundreds of DNA molecule per designed fragments generally stored. Therefore, this factor should also be reflected.

The minimum NGS coverage is the amount of sequencing used to achieve

the criteria for DNA-digital data storage, expressed as a multiple of the total number of fragments. Generally, the larger the NGS coverage, the larger the amount of data, so the error rate can be reduced and the fragment with fewer numbers through uneven amplification can be identified.

	Church *et al.*	Goldman *et al.*	Grass *et al.*	Bornholt *et al.*	Blawat et *al.*	Erlich *and* Zielinski	Organick *et al.*
Error correction method	No	Rep	RS	Rep	RS	Fountain	RS
Information capacity (bit/nt)	0.6	0.19	1.16	0.57	1.18	1.57	0.81
Physical information density (Pbytes/g)	1.28	2.25	25	-	-	214	-
Input data (Megabytes)	0.65	0.75	0.08	0.15	22	2.15	200.2
Full recovery	N	N	Y	N	Y	Y	Y
Number of oligonucleotides	54,898	153,335	4,991	151,000	1 million	72,000	13 million
Minimum NGS coverage (average)	3,000x	51x	372x	40x	160x	10.5x	5x

Table 2.1 Comparision between methods (Rep, Repetiotion method. RS, Reed-Solomon error correction): Full recovery indicates information was recovered. Megabyte : 10^6 bytes, Pbyte : Peta byte, 10^{15} bytes.

2.1.5. Comparison of Cost of DNA-based Data Storage Methods

Previous research has increased the stability of DNA-based data storage methods through various methodologies and enabled the complete storage and restoration of data. However, the field of DNA-based information storage faces problems regarding the practical storage of a large amount of information due to the high cost of synthesizing DNA. At present, the cheapest cost for DNA synthesis reported is 0.05 US dollar per 100 nt[10], and the sequencing cost is 0.0000012 US dollars per 100 nt[24]. and the sequencing price is several ten thousand times cheaper[25]. In addition, with the current development of NGS technology, sequencing prices have been decreasing by 1/10 every year for few years while synthesis rate is much slower (http://www.synthesis.cc/synthesis/2016/03/on_dna_and_transistors), so that the difference between them is likely to increase.

The price of writing and reading in a DNA-based data storage can be obtained by applying the information capacity and minimum NGS average described in Table 2.1 by the DNA synthesis and sequencing prices, respectively. As a result, the most affordable price of DNA-based data storage today is about 3500 US dollars to store 1MB (Figure 2.3)[10]. As shown in the Figure 2.3, even after the data to DNA encoding algorthm was developed and full recovery of the data from the DNA was achieved, there was no dramatic decrease of the cost of the storage.

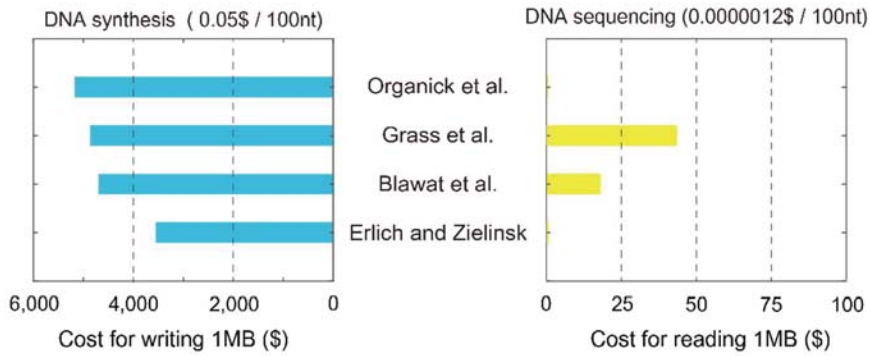


Figure 2.4 Cost comparision between DNA-based data storage methods: Of the proposed studies in the past, only describe the price of recovering data from DNA perfectly. The price was calculated by comparing the number of nucleotides synthesized and the amount of NGS from each study and multipying that with the cheapest synthesizer and analyst in the current market[10], [24]. This figure has been modified from the previous research[17].

2.2. Addition of Encoding characters for Higher Information Capacity

In this dissertation, I propose a method to increase information capacity by adding additional characters other than A, C, G, and T to data to DNA encoding. The ideal additional characters as the restriction that it can be synthesized, amplified and analyzed using existing platforms and the cost should be the same. In that manner, the addition of the encoding characters could increase not only the information capacity, but also cost-based practicality.

Examples of such additional character's candidates include bases whose biochemical properties have been changed from A, C, G or T, such as the methylation modification to the base C or the phosphorothioate addition between the base and the base in addition to the basic base. However, DNA amplification efficiency using an enzyme is not low or modification may be lost during amplification. In addition, it has been reported that unnatural bases such as RNA base, Z, P, dNaM, dm5SICS, isoC, and isoG other than ACGT[26] (Figure 2.6) are synthesized and amplified through an enzyme, which can be used for additional characters. However, these bases are not suitable for use in DNA based storage because they cannot be analyzed by NGS. Also, even if this is possible, the mentioned examples have the disadvantage that the synthesizing cost is expensive compared to the existing ones.

As an example of an additional encoding character, researchers previously

introduced chemical treatment-enabled additional bases[27]. By motivated by the fact that deamination of C to uracil (U) could not applied for the 5-metyl modified C, when treated by the bisulfite ion catalyzed hydrolytic, they use the 5-metyl modification as a new encoding character (Figure 2.5(a)). After the chemical treatment, only un-modified C was changed to T since the base U will be supplemented with A while PCR (Figure 2.5(b)). However, since there are only few types of modification that be used, and the synthetic cost of the modified DNA is more than 10-fold expensive than the normal one, this could not be practically used. From this, researchers tried to use this as a cryptographic system, rather than large-scale data storage system.

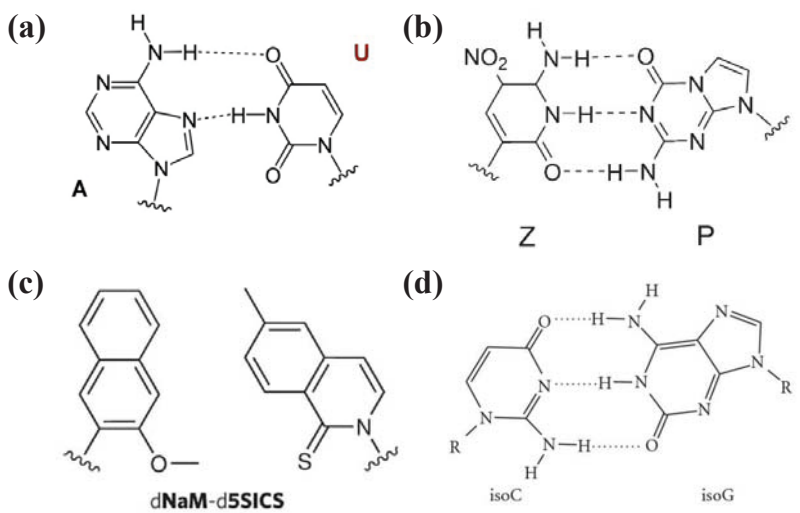


Figure 2.5 Example of unnatural bases and its chemical structure.

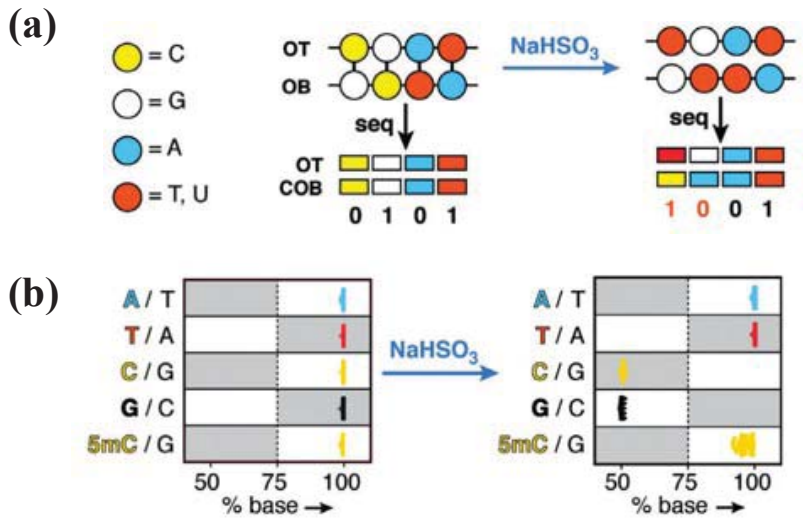


Figure 2.6 A chemical treatment-enabled additional bases for DNA-based data storage: (a) Base C is chaged to T, after $NaHSO_3$ treatmentl; (b) By utilze 5-metyl modified C, which is not affected from NaHSO3, encoding character set for DNA-based data storage has been expanded.

2.2.1. Degenerate Base

In this dissertation, I propose to use degenerate base as a new encoding character. A degenerate base is a mixture of A, C, G, and T at a specific base position. For example, if a particular DNA sequence is ‘CWA’ and W is a degenerate base with mix of A and T, then the DNA molecule of ’CAA’ and ‘CTA’ are presented. There are total of 11 degenerate bases and their names are defined by IUPAC (International Union of Pure and Applied Chemistry) (Figure 2.6 (a))[28]. Also, if the ratio between the bases that make up the

degenerate base could be adjusted, ideally an infinite encoding character can be created.

The chemical DNA synthesis method that is currently used is to make long DNA molecules by sequentially connecting blocks corresponding to a single base[18]. In addition, the method is not a monomolecular synthesis, and it produces 10^9 or more identical molecules even when synthesized at the smallest scale[29]. Therefore, the degenerate base can be synthesized by mixing the block elements corresponding to the base combination in the conventional DNA synthesis process. In the case of sequence analysis with NGS, the number of NGS reads per fragment is increased to confirm the degenerate base, and there is no change in the platform. In the aspect of the cost, since the NGS cost for the DNA is small enough to be ignored, the cost for increased read also would be ignored. This will be discussed in Chapter 5. In addition, in the case of column or ink-jet base DNA synthesis[29], which is most conventional platforms, there is no additional amount of chemical spent and ideally there is no increase in the cost.

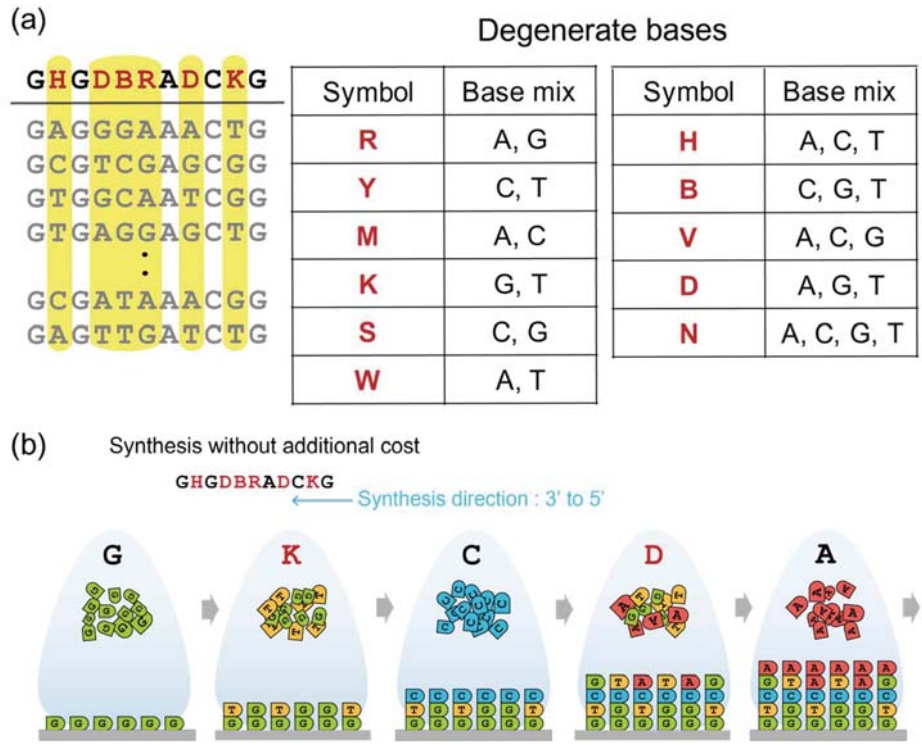


Figure 2.7 Degenerate base and its synthesis: (a) Left : A examples of degenerate base, Right : The symbol of degenerate bases defined by IUPAC[28]; (b) Synthesis skimetic for oligonucleotide including degenerate base. This figure has been modified from the previous research[17].

Chapter 3. Addition of Degenerate Bases to DNA-based Data Storage

In this chapter, a design of the DNA for data storage, experiment approach to handle the DNA for the demonstration of the concept of adding a degenerate base to the DNA-based data storage method will be described. First, this chapter deals with the encoding method of converting digital data into DNA including degenerate bases. Second, chemical synthesis of DNA, its amplification, and sequencing methods will be described. Finally, a method of decoding digital data from sequencing data and its results will be introduced. Through this chapter, hundreds of kb of data was stored in DNA and restored successfully.

3.1. Digital Data to DNA Encoding Method

The DNA codon method was used in this dissertation to convert the binary data consisting of 0,1 into the degenerate base containing DNA. Due to the nature of the degenerate base, the GC contents may change in one base place, so the fountain method to control the GC contents of fragments in the design was not used. In addition, molecular biologic methods, not algorithmic methods, were used to uniformly amplify molecules with various GC contents variants resulting from fragment design, which is covered in Chapter 3.2.

The factor to be considered in generating the DNA codon is how long it will allow homopolymer. Previous studies have shown that increasing the length of the homopolymer beyond 4 nt causes enzyme slippage in the PCR[20]. It is also reported that the error rate of NGS increases when the length of the homopolymer longer than 4 nt or more[21]. Based on this, a codon that allowed only 3 nt homopolymer was generated. First, all kinds of codons consisting of 3 nt were made. After that, the codon that has same base in second and third position of it. If this method is used, even if the codons are arranged successively, only homopolymers having 3 bases or less will be presented. Also, in the case of degenerate bases, if homopolymer was present the oligo was not used, by considering all possible base combinations. In this way, a total of 750 codons were created. This is more than 15 times the codon of 48 without the degenerate base. Finally, encoding was performed by associating two codons

with 19 bits of digital data.

3.1.1. Design of the DNA library for storage

As a pilot study, a DNA fragment of 85 nt was designed, and 40 nt is the adapter size for DNA amplification, so the actual length used for the data to DNA encoding design is 45 nt. The adapter sequence was made by trimming the sequencing primer of the NGS platform (Illumina), that way it could be used both in amplification and sequencing. Sequence of the adaptor is below:

Forward adaptor: ACACGACGCTCTTCCGATCT

Reverse adaptor: AGATCGGAAGAGCACACGTC

42nt, or 14 codons is the length to which data is allocated, and addresses are assigned to the remaining 3nt (Figure 3.1). Addresses corresponded to 48 codons without degenerate base. The reason why the degenerate base is not addressed in the address is because it is necessary to classify the read using the address to find the degenerate base. This will be covered in Section 2.3. In this study, the text file of Figure 3.2 was converted for the pilot test. Its size is 854 bytes and 45 DNA sequences are generated (Table 3.1).

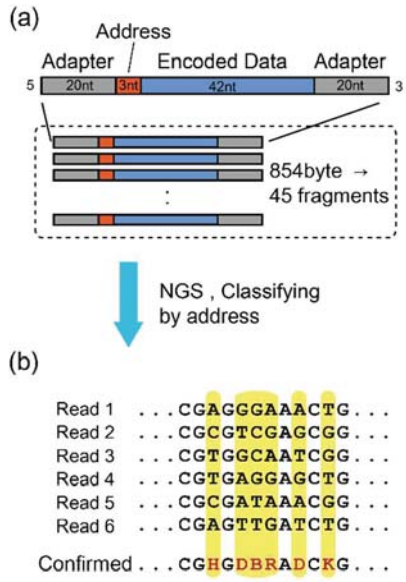


Figure 3.1 Structure of the DNA library. This figure has been modified from the previous research[17].

Figure 3.2 The text file used for encoding in the pilot demonstration. The content of the text file is a member list of the research group, BiNEL (http://binel.snu.ac.kr).

Fragment number	Sequence
1	ACACGACGCTCTTCCGATCTACADYRTKCSATRTCACRRTADTCKGWTWCCYGDWSYTMTTGRAYAGATCGGAAGAGCACACGTC
2	ACACGACGCTCTTCCGATCTCCAYCWNWCMWSWGYMSAGMGBTGRWCKASSDCTYABTMYKCGCDAGATCGGAAGAGCACACGTC
3	ACACGACGCTCTTCCGATCTTCAAAKCABNGYYYACAGCTVWYGDGWRTRRTCNMTKCRNCGRGMAGATCGGAAGAGCACACGTC
4	ACACGACGCTCTTCCGATCTGCAKSWDSWNGABKMYWCVCDHTRYHGGCTDCWTGHNSADCACCGAGATCGGAAGAGCACACGTC
5	ACACGACGCTCTTCCGATCTATAWSWWTAAWSDMGHGMNAGCTMCYGWMTMWCRCGYCGNKMYTAAGATCGGAAGAGCACACGTC
6	ACACGACGCTCTTCCGATCTCTAMTRYCWRTSNAYYKCTAYWKMNWSBTSVMTWWSTDCCTGVYGAGATCGGAAGAGCACACGTC
7	ACACGACGCTCTTCCGATCTTTASHGHGMNTSAWSKGCWDCAAKTTMBASAYACKCTYACTGKGHAGATCGGAAGAGCACACGTC
8	ACACGACGCTCTTCCGATCTGTABGWKCKGGCMTVSGABWGATMDABTTRCCDGGADGAWVTYYAAGATCGGAAGAGCACACGTC
9	ACACGACGCTCTTCCGATCTAGABKARCGHWSDWCTTCWACACKCMKVGWGRYRGTGGTACRAYRAGATCGGAAGAGCACACGTC
10	ACACGACGCTCTTCCGATCTCGABKAMAGSWCBRCCRTNDCVTMYRTVDCDGATAYDSANWCDASAGATCGGAAGAGCACACGTC
11	ACACGACGCTCTTCCGATCTTGADCKHRYHGYCMGNSWVTSBTGRTMKASBSARCGVSTVCAAABAGATCGGAAGAGCACACGTC
12	ACACGACGCTCTTCCGATCTGGAVMKHTCRBABTACASHWSVMTSYRYASHTRKWSTWSBCABKMAGATCGGAAGAGCACACGTC
13	ACACGACGCTCTTCCGATCTAACCAKMGMNTSCGWRAKSTAGGYTSTCGWTCKGGHTMGATGGCDAGATCGGAAGAGCACACGTC
14	ACACGACGCTCTTCCGATCTCACRYGAAYHCTRGCKCRMCDCTMTCTVRTBTACYADATATGNGYAGATCGGAAGAGCACACGTC
15	ACACGACGCTCTTCCGATCTTACDKCDGTVAGSTRGGHBGTKATVYAVMTTRCCKARSTSAGRYAAGATCGGAAGAGCACACGTC
16	ACACGACGCTCTTCCGATCTGACTCWVBAWGYVYRCASGABBTGDGTBSWDTGMYAGCKMSTCHGAGATCGGAAGAGCACACGTC
17	ACACGACGCTCTTCCGATCTATCKYRVAYKTSMMTRCTVHGHTRDWCVMTTRCVYAKATNBAYCKAGATCGGAAGAGCACACGTC
18	ACACGACGCTCTTCCGATCTCTCAMKSAKWTSHCWHGMRSTMRYMYAGTRHRTYSATTGWKCBMTAGATCGGAAGAGCACACGTC
19	ACACGACGCTCTTCCGATCTTTCTCWDBAGGCNTVHGMGVTTMKDGHVRTYKCCKAACTCTGCRTAGATCGGAAGAGCACACGTC
20	ACACGACGCTCTTCCGATCTGTCDKCNMTDVTMCAHCGTAYHRYGKCRSTSGABYGNAKBAKMWCAGATCGGAAGAGCACACGTC
21	ACACGACGCTCTTCCGATCTAGCDKCVTRAWSVCRHTAKWGVTMAGWYAYKMTKGTHTAATRVTCAGATCGGAAGAGCACACGTC
22	ACACGACGCTCTTCCGATCTCGCWAKMWSKRYKGCHGMKVTKYRRATTCGGGHMKMDWGYATKDCAGATCGGAAGAGCACACGTC
23	ACACGACGCTCTTCCGATCTTGCDKCDGTVAGVYGCCGNAKBCKVRYYSWKTSBYGHAKKDCMGCAGATCGGAAGAGCACACGTC

24	ACACGACGCTCTTCCGATCTGGCHCWACDRWGBTADGTRGAHTGHKMK ASKMTHWSWCRWSAHAGAGATCGGAAGAGCACACGTC
25	ACACGACGCTCTTCCGATCTAATYCKMGTTTCYGMTASWSABCKTTMKM TNABBYGNAKBAKMWCAGATCGGAAGAGCACACGTC
26	ACACGACGCTCTTCCGATCTCATTCKRKMNBAWSWWCTNHGVMTSYRD MTHYGBRYKTASMTVMGAGATCGGAAGAGCACACGTC
27	ACACGACGCTCTTCCGATCTTATCGTSTMTWSTTAAKCGSWKAYKCGYS WAWSVACSTRATRBTCAGATCGGAAGAGCACACGTC
28	ACACGACGCTCTTCCGATCTGATWCKGACNBAWSWWCTNHGDCTYGW RSWKRCRCGYVTVRTGGMAGATCGGAAGAGCACACGTC
29	ACACGACGCTCTTCCGATCTACTWCKSTCTWSYTASTAAWSGMKDCTDS ABGHVWSBCTTRTSTMAGATCGGAAGAGCACACGTC
30	ACACGACGCTCTTCCGATCTCCTTCWDBAMTSBAKHTAWASGMKGKCG DCNCKSYRAACSTMGACAGATCGGAAGAGCACACGTC
31	ACACGACGCTCTTCCGATCTTCTWAKBYAKYGYSTYSAMASYRYVTVDR YTGCGWCHTMADCKCRAGATCGGAAGAGCACACGTC
32	ACACGACGCTCTTCCGATCTGCTYCWSACVBAYKCCASHWSAGAKGCYA YKMTARTRKCMKCARYAGATCGGAAGAGCACACGTC
33	ACACGACGCTCTTCCGATCTAGTTCWVBAWTSHCWHGMKCGDTMTAGD MTBYGDWSHTANCAHCTAGATCGGAAGAGCACACGTC
34	ACACGACGCTCTTCCGATCTCGTYCWVACNGAKGYTRTRCDKYRDKCSD CTTRSMKNGWNCARTGAGATCGGAAGAGCACACGTC
35	ACACGACGCTCTTCCGATCTTGTCAKMYATWSVCADTAHCTVCKDRYSD CTTRHMKYTCVRTBTMAGATCGGAAGAGCACACGTC
36	ACACGACGCTCTTCCGATCTGGTTCKKAKHCTGGCTATRABCTRTSTBTS VMTGSWGGHMTCTWGAGATCGGAAGAGCACACGTC
37	ACACGACGCTCTTCCGATCTAAGDKCVGTBAGVGABTMHMGHRYCKCV RTMKCBYGNAKGTGDBAAGATCGGAAGAGCACACGTC
38	ACACGACGCTCTTCCGATCTCAGTCWDBAGGCNTVHGMKVTKYRAKMK ASCSAMYANTAWAGSTGAGATCGGAAGAGCACACGTC
39	ACACGACGCTCTTCCGATCTTAGVKCBABBCGGKCKGWHMKKYRATCD SAVTVMRCRWSBAKMWCAGATCGGAAGAGCACACGTC
40	ACACGACGCTCTTCCGATCTGAGWAKYWSCWSSABHGMVSTGMKKCTC DCWGATTMHTCGRTGMTAGATCGGAAGAGCACACGTC
41	ACACGACGCTCTTCCGATCTACGCGTWTMVASHSWTVTKMTWGYGABN SAATAVWSTCTVRTDGMAGATCGGAAGAGCACACGTC
42	ACACGACGCTCTTCCGATCTCCGSGMHKATRCHKMCAGNABHCWSATAT CGSWWSANGCGRTCDCAGATCGGAAGAGCACACGTC
43	ACACGACGCTCTTCCGATCTTCGSHGVSAHCTRGCMKCBKARKAWRYDS AMGWYSATTGHRCKTGAGATCGGAAGAGCACACGTC
44	ACACGACGCTCTTCCGATCTGCGNAYMWSHTSYKMGGABVTWAYWCGY ATYYABYAAHGVCAAABAGATCGGAAGAGCACACGTC
45	ACACGACGCTCTTCCGATCTATGVMKHTCSACNCRCASMABRKAWRYDS AMGWCSAVGTASTVTMAGATCGGAAGAGCACACGTC

Table 3.1 Sequence list of the designed DNA library for data storage.

3.2. Amplification and Sequencing of DNA library

The DNA corresponding to the designed fragment was purchased from the Macrogen (Seoul, South Korea), by using a column-based synthesizer. From this, there was no cost increase for the DNA according to the provider, when compared to the non-degenerate base only sequence. The synthesized DNA was collected into a tube and diluted to a concentration of 800 molecules per fragment at 1 ul. In the case of library amplification, the KAPA library preparation kit from the KAPA bioscience was used, by following previous studies have shown that KAPA library preparation kit is suitable for amplifying various libraries of GC content[30], [31]. In addition, all amplification procedures were performed using qPCR to prevent excessive amplification. Sequences of the forward and reverse primer are:

Multiplexing Read 1 Sequencing Primer (Forward)

ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Multiplexing Read 2 Sequencing Primer (Reverse)

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

The amplified samples were then analyzed using the Illumina Miniseq with a 300cycle pair-end read protocol.

3.3. Decoding of the Data from the Sequencing Data

The raw data resulting from the sequencing is the fastq file of the pair-end reads. The PEAR algorithm was used to stitch this pair-end reads. After that, NGS reads with the appropriate lengths, same as the designed fragment since one could not know the place of the insertion or deletion and the read with different length could not be properly aligned. There could be the clustering method to solve the problem and this will be discussed in Chapter 5.4.1.

Finally, the duplicated reads were deleted. This can eliminate the bias effect of amplifying only specific fragments. The number of reads and its ratio to the raw data is described in Table 3.2. Here, the amount of heterogeneous read is dramatically smaller than that of raw data. This is because the sequencing data is excessively large, resulting in a redundant read. It can be inferred that recovery of data may be possible by further reducing the amount of data, which is covered in Chapter 3.3.2.

Before Assemble	162707	100%
Assemble	158260	97%
Length filter	127082	78%
Heterogeneous reads	26675	16%

Table 3.2 Number of read and its ratio to the raw data, that acquired from each step. This table has been modified from the previous research[17].

3.3.1. Determination of Degenerate Base

To determine the degenerate base in sequencing data, there must be a criterion that can be used to determine whether the ratio of base calls is error or degenerate. For this, the process of obtaining the criterion for determining the degenerate base by checking the distribution of ratio of A-C-G-T in each aligned position such as yellow bar of Figure 3.1 (b) has been proceeded. Figure 3.3 is the scatter plot of the normalized ratio. 15 clusters for 15 characters including 4 bases and 11 degenerate bases could be decided by k-mean clustering algorithm or heuristically. The determined characters were identical to the designed base without error.

However, If the type of degenerate base used in encoding is not known or when there is a large amount of data, the use of clustering algorithms may be limited or slow. For this reason, a process to simplifying determination process is proposed. Proposed method performs clustering to check whether the base call in the ratio is an error or an intended from the design. First, draw a histogram of the ratios of each base. The leftmost part with a very small ratio less than 0.1 in the histogram can be judged as an error (Figure 3.4). To separate it the first inflection point of a graph was find as a decision line. From this, the left side of this decision line was considered as not intended base for design and the right side was considered as the intended base for design. Finally, the base could be determined by combining of the intended base. For example, in a

position, if ratio of A, G and T is over its decision line and C is not, the determined base will be the D, according to Figure 2.6 (a).

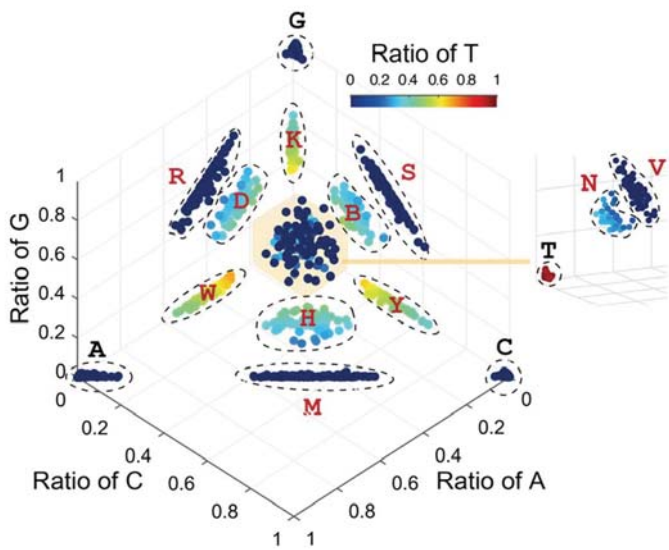


Figure 3.3 Scatter plot of the ratio of bases in the same position. Degenerate base could be determined. This figure has been modified from the previous research[17].

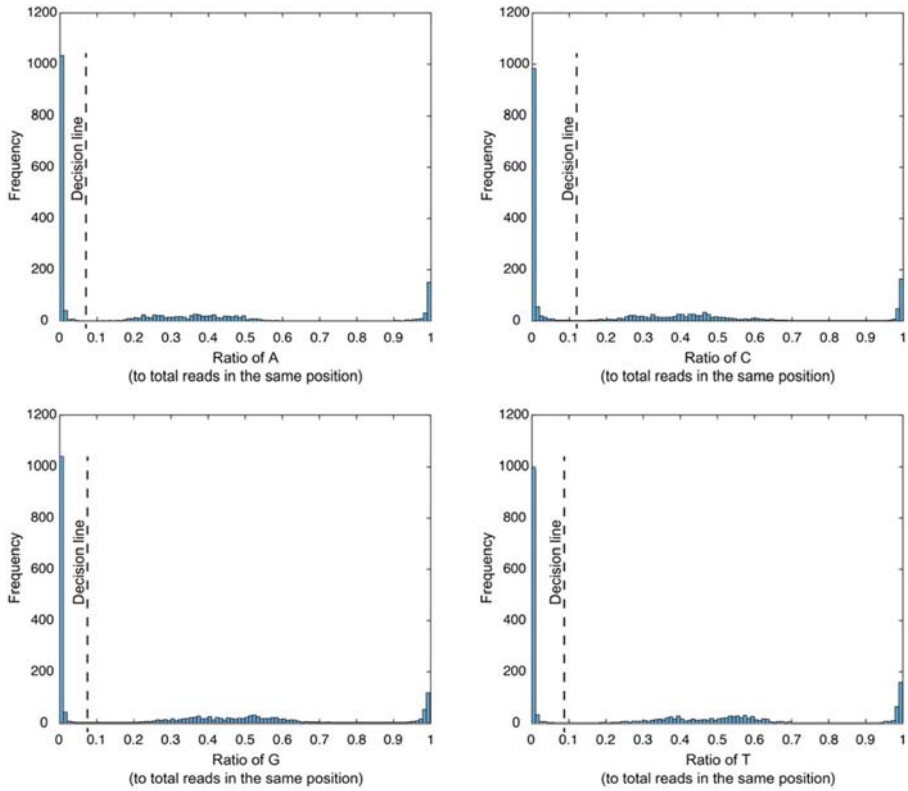


Figure 3.4 The histogram of the ratio of base in a position in the sequence. Decision line could be determined. This figure has been modified from the previous research[17].

3.3.2. Decoding Result and Down-sampling of Sequencing Data

By comparing the DNA sequence after the degenerate base determination with the original design, the error rate of this platform was calculated. The sequence extracted using raw NGS data was the same as the design, so there was no error. Then, NGS raw data was reduced by random down sampling to judged how small data could be extracted without error. NGS depth is the number of NGS read divided by the total number of fragments designed. The number of reads from the raw data is 162707, which has a depth of 3600x. Ideally, if there is no degenerate base and there is no error, all data must be read with a depth of 1x. However, it has been reported that data can be recovered with error correction by minimum depth of 5x as in Table2.1. Figure 3.5 shows the error rate according to NGS depth. The mean and standard deviation were obtained through five random sampling. As shown in this figure, at least 250x depth was enough to restore data without error. Also, it can be seen in Table 3.3 that the amount of duplicate readings decreased through down sampling. In other words, it can be seen that the depth of the NGS was excessively higher than the number of molecules in the sample. Also, in this experiment, since there is no error correction, it is not possible to correct a single error that occurred after 250x, which can be solved by adding an error correction in the next chapter.

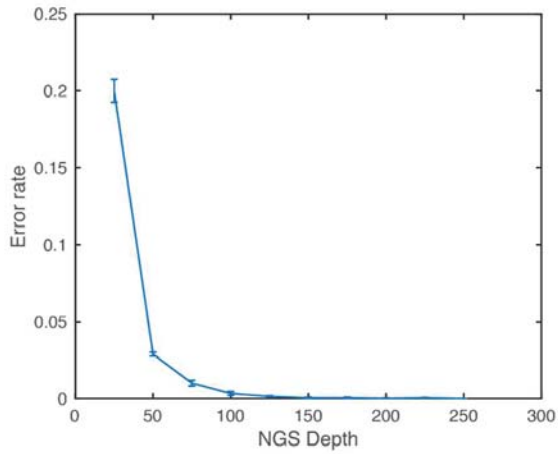


Figure 3.5 Error rate due to sequencing coverage. The standard deviation (s.d.) of the experimental results were obtained by repeating the random sampling 5 times. The error bars represent the s.d. This figure has been modified from the previous research[17].

	3500x		250x	
Before Assemble	162707	100%	11250	100%
Assemble	158260	97%	10949 (12.63)	97%
Length filter	127082	78%	8808.2 (31.73)	78%
Heterogeneous reads	26675	16%	7149.8 (23.27)	64%

Table 3.3 Number of read and its ratio to the unfiltered data, that acquired from each step. 3500x is the raw data, and 250x is the data obtained through five random downsampling. The parentheses are the standard deviation. This table has been modified from the previous research[17].

Figure 3.6 is a box plot of the GC contents that can be observed in each fragment design due to the degenerate bases. Figure 3.6 (a) shows the GC contents that can be observed in the design and 3.6 (b) is the GC contents observed in the experimental results. Since the sampling size (~592x, according to the number of heterogeneous reads) is significantly smaller than the almost 10^9 variables that can be derived from the designed fragment, quantitative analysis was not possible. However, averages of GC contents from the experimental result were found to be similar to the design. To solve this problem, read number of all individual sequences was normalized, by the read value when assuming the even representation of sequence in the designed fragment. After that, the average representation of the sequences in each GC contents was plotted (Figure 3.7(a)). From the figure, when the GC contents were between 30 and 60%, there was no uneven representation, but in out of the range, uneven amplification was confirmed. Also, there is dramatically high representation of few fragments with GC contents around 80%, which tendancy is different from the data of previous research[19]. However, this does not seem to be meaningful because the amount of data is less than 10. But even so, more than half of the sequences within the designed fragment are within the range, so decoding was possible. A simulation approach to this is discussed in Chapter 4.

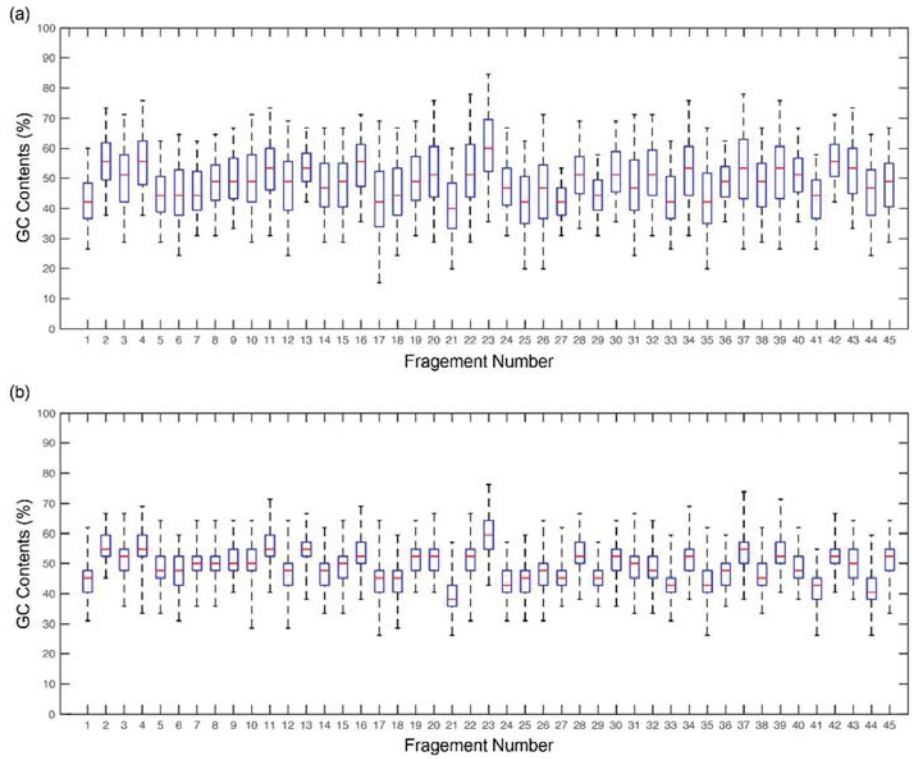


Figure 3.6 Box plot of GC contents variants due to the degenerate bases from each fragment designed: (a) GC contents from the design; (b) GC contents from the experiments.

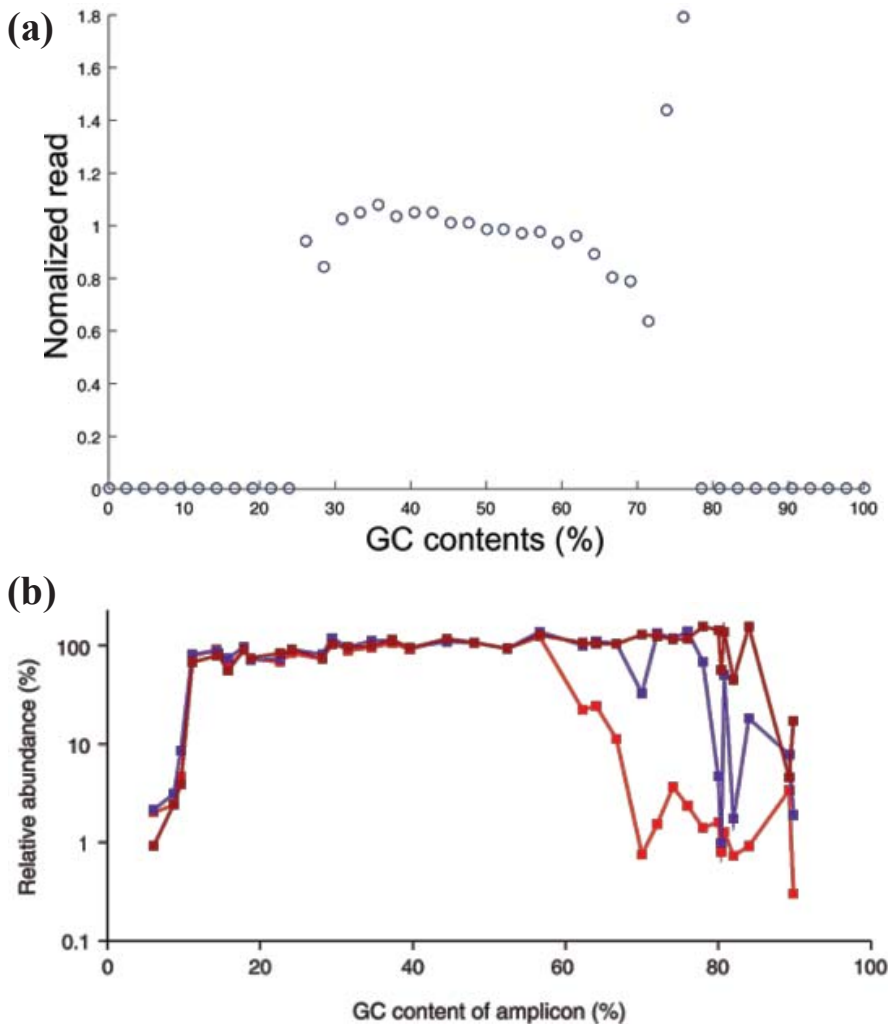


Figure 3.7 (a) Nomalized read number of the sequences from the NGS data, according to GC contents. The closer the value is to 1, the smaller the effect of the uneven representation. (b) Experimental data follows the tendancy of previous research. The figure has been modified from the previous research [19].

The results obtained through this pilot experiments are summarized in Table 3.4. When comparing this study with the most efficient study reported, information capacity and physical information density have more than doubled. However, the size of the encoded data is very small when compared with the existing studies, and the scalability of the platform cannot be confirmed.

	Error correction method	Information capacity (bit/nt)	Physical information density (Pbytes/g)	Input data	Full recovery	Number of oligonucleotides	Minimum NGS coverage(average)
Erlich *and* Zielinski	Fountain	1.57	214	2.15 (Megabytes)	Y	72,000	10.5x
This research	N	3.37	485	854 (bytes)	Y	45	250x

Table 3.4 Summary of the result. The result is compared with Erlich and Zielinski[10], that achieved both highest information capacity and physical information density. This table has been modified from the previous research[17].

3.4. Microarray-derived DNA Pool Based DNA-based Data Storage

To supplement the lack of scalability as mentioned in the results of Chapter 3.3, a new experiment was conducted. To do this, hundreds of thousands of DNA must be synthesized. In addition, there should be the method, or algorithm to handle errors and losses that may occur during handling of many types of DNA. First, the column base synthesis that used for the demonstration is a method to synthesize tens of nano moles of molecules for one designed fragment. However, this synthesis method is not suitable for synthesizing large amounts of DNA because the synthesis price is $0.05-0.15 per nucleotide. To solve this, the synthesis method had been changed to a DNA-microarray-based synthesis method, which is about 100 times lower than the column-based synthesis method. However, this method has a drawback that the error rate is higher than that of the conventional method. Also, there may be a loss of fragment due to uneven amplification that occurs when a large number of designed fragments is contained in a pool and amplified. In order to solve problems mentioned above, error correction was included in the design aspect.

3.4.1. Design and experiment of the DNA library for storage

For the microarray-derived DNA oligopool synthesis, B3 Synthesizer DNA microarray synthesizer (Customarray Inc. USA) was used. Due to the limitations of the synthesizer, there were design limitations in this demonstration. First, the synthesizer was unable to mix the phosphoramidite during synthesis. So, to use degenerate base, a line of mixed phosphoramidites should be added. Since the synthesis platform had only six lines, we only used two degenerate bases. Also, maximum synthesis length was about 160nt, which is shorter than DNA library that other reports used.

From this, I encoded the thumbnail image of Hunminjeongum Manuscript (or Hunminjeongum Haerye, Figure 3.8), which is the UNESCO memory of the world registered documented heritage submitted by Republic of Korea in 1997. Image file was resized to 692 × 574 and the file size was 135,393 bytes. I used W (mix of A, T) and S (mix of C, G). Each 7-bit data is matched to a DNA codon consisting of three combinations of alphabets from six. The three-nucleotide codon is designed so that the whole sequence does not contain more than three repeating nucleotides, including those in the degenerative parts, same as chapter 3.1.1. The encoded information is divided into fragments of 111 nt, and an address composed of three address DNA codons, total length of 9 nt is assigned thereto. Each fragment is supplemented with an adapter in chapter 231.1 for amplification and sequencing.

4183 fragments in total were generated from the data fragmentizing. After that, I also add Reed-Solomon based error correction (Figure 3.9). I designed 9 redundancy fragments in every 118 information fragments to correct 4 false information or 9 missing information in maximum. Also, 5 redundancy fragments were added in block of 53 fragments to correct 2 false information or 5 missing information in maximum. Finally, 4503 DNA fragments can be designed.

Designed DNA library was synthesized using 12k microarray following standard protocol provided (Customarray Inc. USA). After that, the synthesized library was quantitated using qPCR. Relative sample quantification was accomplished by interpolation from a standard curve, generated from DNA samples of known concentration. From this, the synthesized DNA library has the 438 molecules per fragment, in 1ul of sample (1974204 molecules, standard deviation: 81696). A sample of 1ul was treated in the same manner as in Chapter 3.2.

Figure 3.8 The thumbnail image of Hunminjeongum Manuscript (or Hunminjeongum Haerye), which used for encoding. The size of file is 135,393 bytes.

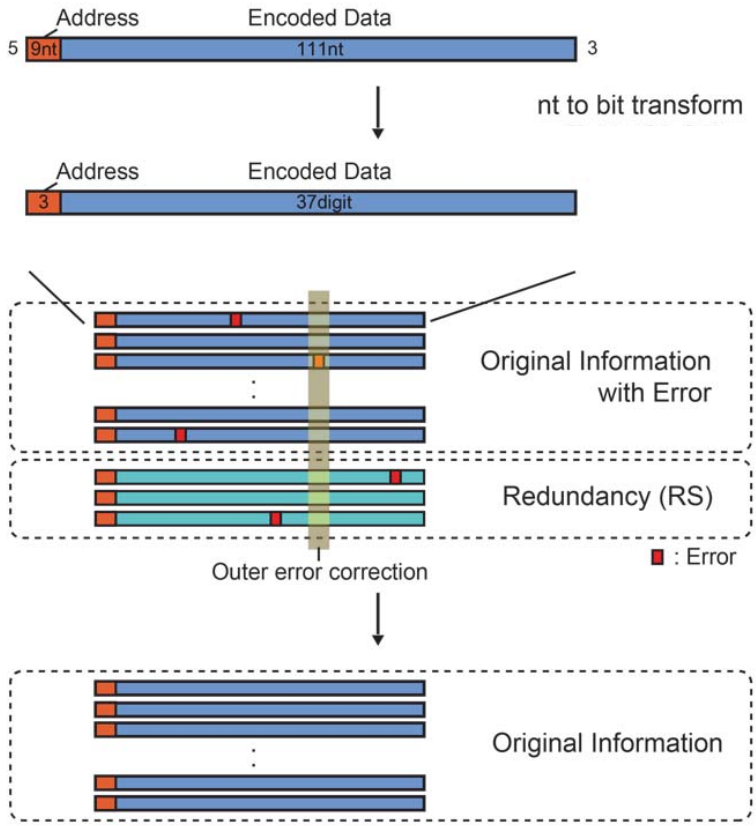


Figure 3.9 Stucture of the DNA library. Error could be corrected with the Reed-Solomon (RS) based redundancy. This figure has been modified from the previous research[17].

3.4.2. Experimental Result and PCR bias analysis

The process for determinizing the degenerate base is the same as Chapter 3.3.1. Figure 3.10 shows the clustering results of W and S in the A-T or G-C domain. Since the coupling efficiency during synthesis varies for each base, by type and position in the growing oligonucleotide, the intermediate ratio of the nucleotides analyzed was not equivalent, or their average was not 0.5. The information was successfully recovered from raw data with an NGS depth of about 1250x. The results of filtered reads in each step for decoding are shown in table 3.5. Unlike the initial demonstration, the number of reads is reduced by about half when filtering the length of the appropriate DNA, which is the effect of deletion during DNA synthesis. Since twice the length of the DNA was used when compared with the initial demonstration, indicating that the deletion errors resulting from synthesis have accumulated.

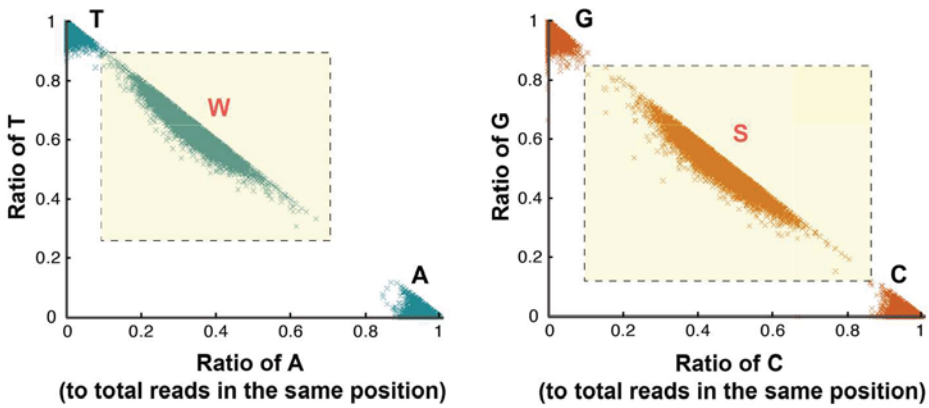


Figure 3.10 Scatter plot of the ratio of bases in the same position. Domain was limited to A-T or C-G to determine the degenerate base, W or S. This figure has been modified from the previous research[17].

Before Assemble	5847136	100%
Assemble	5660429	97%
Length filter	2928269	50%
Heterogeneous reads	1083343	19%

Table 3.5 Number of read and its ratio to the unfiltered data, that acquired from each step. This table has been modified from the previous research[17].

In order to determine the minimum amount of NGS data that is most efficient in recovering the data, recovery of the data through down sampling has been carried out. Figure 3.11 shows error rate of sequenced base pairs in fragments of sampled heterogeneous read depth when the respective decision boundary is applied. The read depth of 25 or more shows error rate of 0.5% or less, which is lower than the combination of the synthesis error and the sequencing error when the normal microarray-derived DNA oligopool is used. However, due to the synthesis and amplification bias, each fragment had uneven read depth, even if the NGS coverage was increased. From this reason, the average error rate of the synthesized library according to NGS coverage was converged to ~0.07% in 500x (Figure 3.12). I also have recovered the data in 10 cases out of 10 random down-sampling the NGS coverage to 250x.

Also, when the number of filtered leads per sampling step is checked in Figure 3.13, It can be seen that the number of heterogeneous reads converges to about 200x, which means that the amount of data available is limited since 400 per molecule fraction of the sample used in the experiment.

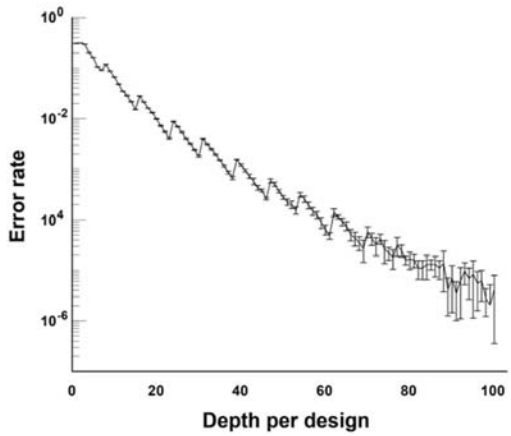


Figure 3.11 Error rate of sequenced base pairs in fragments of specific heterogeneous read depth. The standard deviations(s.d.) were obtained by repeating the random sampling 5 times. The error bars represent the s.d. This figure has been modified from the previous research[17].

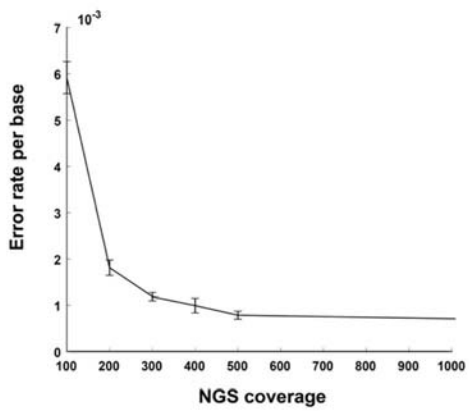


Figure 3.12 Error rate due to sequencing coverage. The standard deviation (s.d.) of the experimental results were obtained by repeating the random sampling 5 times. The error bars represent the s.d. This figure has been modified from the

previous research[17].

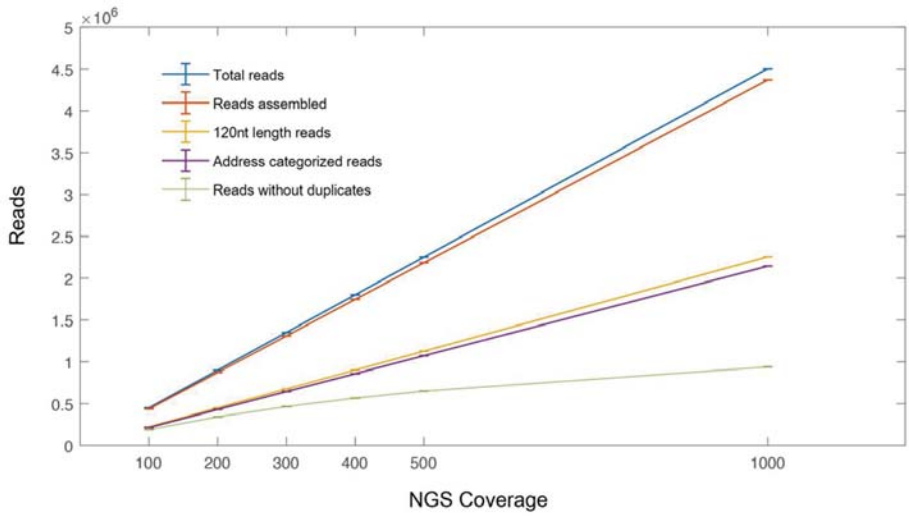


Figure 3.13 Number of read that acquired from each step, according to the NGS coverage of the raw NGS data. The standard deviation (s.d.) of the experimental results were obtained by repeating the random sampling 5 times. The error bars represent the s.d. This figure has been modified from the previous research[17].

To determine the tendency for uneven amplification between fragments, the number of NGS reads actually used for decoding in each fragment was obtained and its probability density histogram is shown in figure 3.14. Past studies have estimated the distribution at other depths by fitting the distribution of read obtained at a specific depth to a negative binomial function[10]. Even though each histogram plot is a match to a negative binomial plot, which follows equation:

$$y = f(x|r,p) = \binom{r+x-1}{x} p^x (1-p)^x$$

However, the approach seems to be not possible, as the number of heterogeneous readings and the associated histogram mean tend to be saturated for depths greater than 500x.

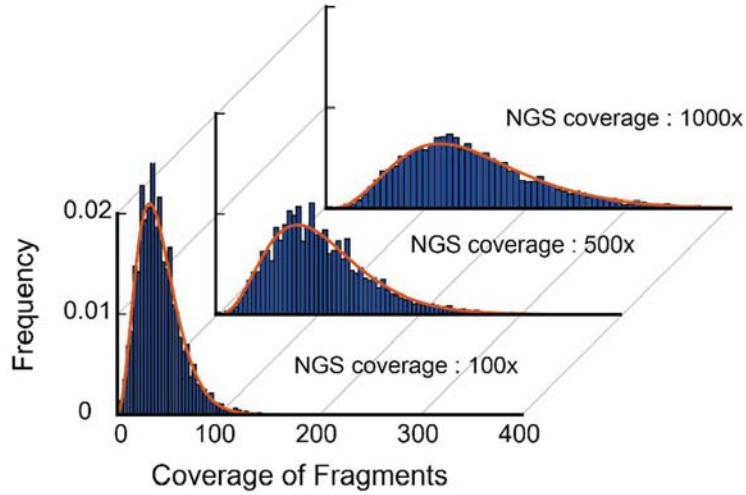


Figure 3.14 Profile of uneven representation of fragments could be seen in the probability density histogram. Red, negative binomial fit. This figure has been modified from the previous research[17].

Finally, a summary of the overall experiment results is shown in Table 3.6. This experiment shows that the information capacity and physical information density increase by more than twice as much as the highest reported in past, while storing up to several hundred kilobytes of data. In particular, in the previous studies, it was mentioned that thousands of molecules per designed fragment were required for storage. However, through this experiment, it could be shown that information can be recovered even with a much smaller number of molecules.

Number of encoding Charaters	15	6	
Input data	854byte	135Kbyte	
Number of oligonucleotides	45	4503	
Minimum NGS coverage (average)	250	250	Highest reported
Net information capacity (bit/nt)	3.37	2	1.55
Physical density (Pbytes/g)	772	485	214

Table 3.6 Summary of the result. The result is compared with Erlich and Zielinski[10], that achieved both highest information capacity and physical information density. This figure has been modified from the previous research[17].

Chapter 4. Simulation Approach for Error Rate Analysis and Cost Projection of Platform in Scaled-up Data Storage

In this chapter, a simulation approach to the concept of adding a degenerate base to the DNA-based data storage method is described. The error rate of the platform in terms of NGS coverage for data recovery when various types of degenerate bases are simulated, in a large scale of the data.

4.1. Monte-Carlo Simulation for Error Rate Analysis

In addition to the experimental results, the error rate of the platform in terms of NGS coverage for data recovery when various types of degenerate bases are simulated. Monte Carlo methods are used for the simulation, which is used to obtain numerical results with repetitive random sampling. For this, in this chapter, random generation of data and its encoding, and random generation of NGS data are performed. To do this, it is necessary to determine what probability distribution the sequencing data followed. Through the results of Chapter 3, call frequency of each fragment designed was analyzed and confirmed that it follows the negative binomial distribution. However, the distribution of bases randomly called from a specific degenerate base position was not confirmed. To do this, the base call at the degenerate base location in the NGS data in Chapter 2.4 was analyzed. First, the fragments with more than 50 read calls were selected and 50 reads from every fragment were randomly sampled. A histogram of the elements that form the degenerate base was plotted (Figure 4.1). From this, the probability distribution follows the binomial distribution (red line), which is the equation:

$$P(x) = p^x(1-p)^{n-x}\binom{n}{x}, n = 50$$

This can be considered to be the probability that the blocks corresponding

to each base are binomially connected during DNA synthesis. Also, from the fitted distribution, we could extract value p, which was used from the simulation. As mentioned in the Chapter 3, the average of the base occurrence comprising degenerate base was not equal, due to the coupling efficiency is different between each base[32].

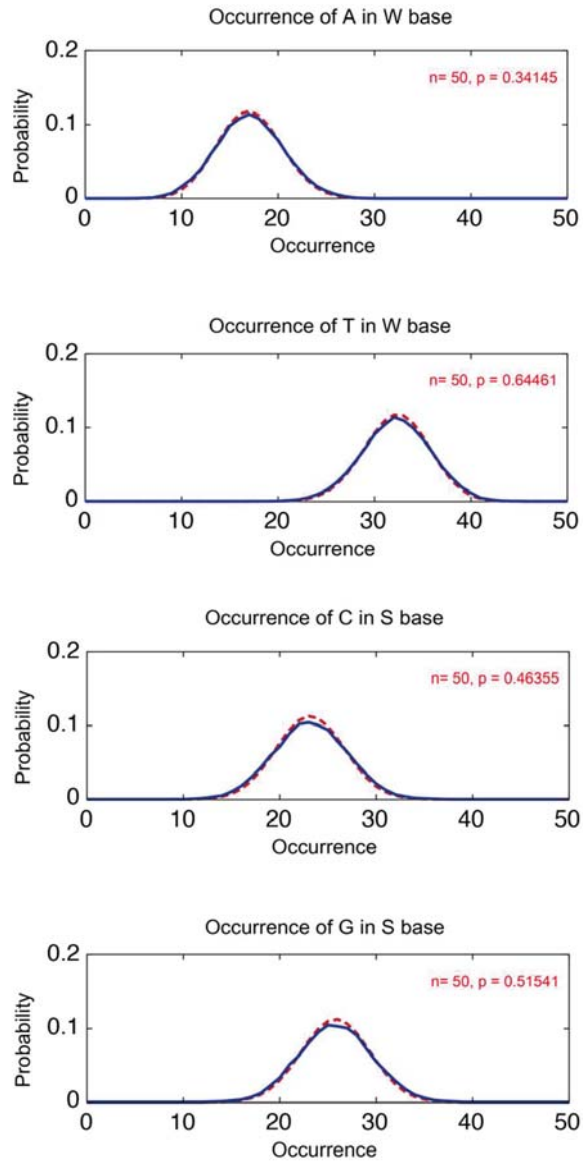


Figure 4.1 Occurrence of base calls that comprises a degenerate base. Blue: histogram, Red: Fitted binomial graph. This figure has been modified from the previous research[17].

The platform was modeled using Monte Carlo simulation. First, simulation using the data that used in Chapter 3 was proceeded, to see if the simulation results are reliable. First, we generated the sequencing results, when assuming there is no uneven representation of the data. The fragment encoded in the experiment was used as an input. After that, sequencing results for the determined number of reads was generated, using the random generation following the binomial probability distribution. Also, error base was generated, following the binomial distribution and p= 2%. After generation of the data, the error rate was confirmed. The result is described in Figure 4.2. Simulation results were similar when compared to the error rate of even sampling in actual experimental data. After that, the uneven distribution (Figure 3.14) obtained in the experiment was applied in the simulation. As a result, an error rate similar to that of the actual test results was confirmed (Figure 4.3). Based on these results, a simulation using an additional degenerate base was proceeded.

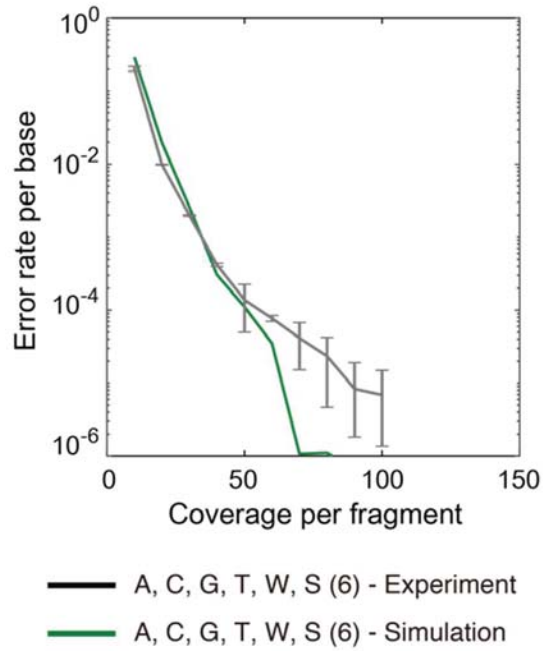


Figure 4.2 The error rate per base pairs according to read coverage of fragments, on which the reads were randomly and uniformly generated in simulation. The experiment data is from Figure 3.11. This figure has been modified from the previous research[17].

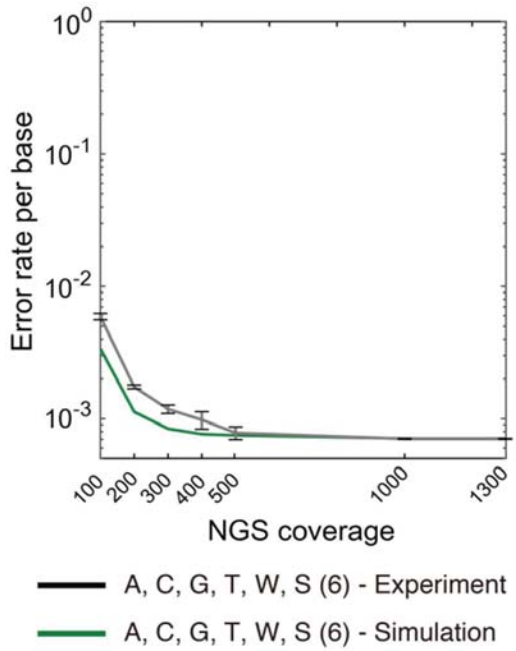


Figure 4.3 The error rate per base pairs according to read coverage of fragments, when applying uneven representaion profile applied. The experiment data is from Figure 3.12. This figure has been modified from the previous research[17].

Based on the data, the error rate per base pair of the models by using 10 various sets of degenerate bases, or all degenerate base was simulated. First, data was encoded after the generation of random data correspond to the one fragment. The length of the fragment used in the simulation was 200nt with a 20nt adaptor that used in the experiment at both ends. 12nt was assigned to the address and the data was stored in 148nt. The encoding procedure was same as described in Chapter 3. After the encoding, the NGS data was generated following uneven representation of the fragments, by applying the probability density of Figure 3.14. In the sequencing results, the base corresponding to the degenerate base are generated with binomial and the mutual probability was assumed as the same. Also, error base was generated, following the binomial distribution and p= 2%. In order to reflect the difference of amplification efficiency according to GC contents, only the read with the GC contents between 40% and 60% was generated. Decoding was proceeded as described in Chapter 3 and error value was obtained.

In the simulation, the extended degenerate bases sets, which were specified by two nucleotides with different ratios (*e.g.*, W1 for A:T=3:7 and W2 for A:T=7:3) was also introduced and expanded the number of encoding characters to 21. For determining the extended degenerate base, the decision was proceeded by comparing the ratio between the two bases. For example, if the base is classified as W, if the ratio of T is bigger than W, it considered as W1. In all simulation, the error rate was obtained by repeated encoding and

decoding of tens of gigabytes.

Figure 4.4 is the error rate of the system, when there is no uneven representation between the designed fragments. Also, Figure 4.5 is the error rate with the uneven representation of the fragments. The use of various types of degenerate bases increases the error rate but the trend of decreasing error rate is shown with increasing NGS coverage.

As an additional experiment, random 100MB of the data was generated and decoded for ten times. In NGS depth of 1300x or more, decoding of 100 MB with 10% error correction of Reed-Solomon proceeded without error. From this, information capacity of 2.67 and 3.05 bit/nt were achieved, when using 15 and 21 encoding characters.

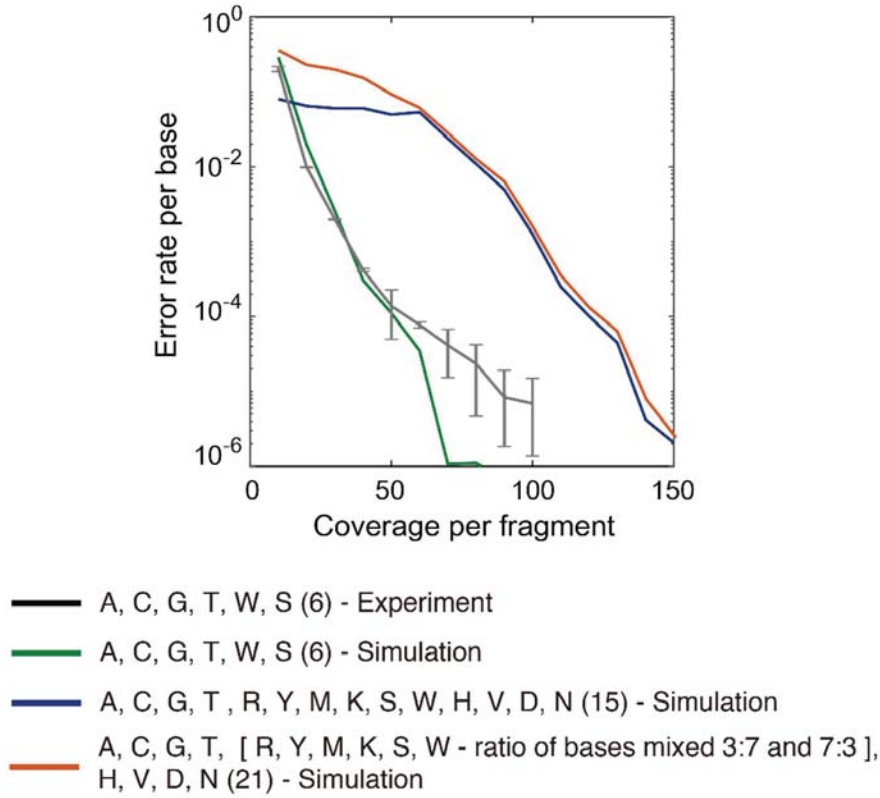


Figure 4.4 The error rate per base pairs according to read coverage of fragments, on which the reads were randomly and uniformly generated in simulation. The experiment data is from Figure 3.11. This figure has been modified from the previous research[17].

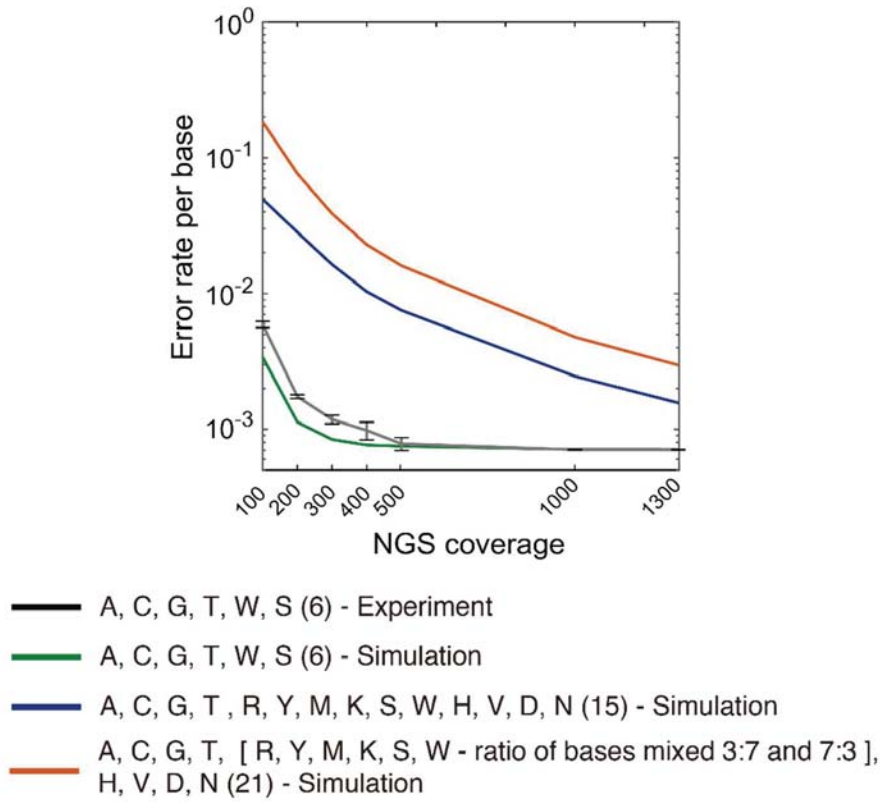


Figure 4.5 The error rate per base pairs according to read coverage of fragments, when applying uneven representaion profile applied. The experiment data is from Figure 3.12. This figure has been modified from the previous research[17].

Chapter 5. Conclusion and Discussion

In this chapter, the information capacity and cost achieved through experiments and simulations in this dissertation will be discussed. In addition to this, how this method can be applied to existing DNA-based data storage algorithms to increase efficiency is discussed. Finally, future works that can develop the proposed method or topics on other problems of DNA-based data storage will be presented.

5.1. Comparison of the Result with Previous Works

From the result of Chapter 3 and 4, we achieved highest information capacity, when compared to the previous researches (Figure 5.1, Table 5.1). I achieved 3.38 bit/nt experimentally, which is higher than the simulation result of 3.05bit/nt, because there is no error correction code assuming a large-scale experiment and the number of nucleotides assigned to the address is short. Even though the result from the simulation is lower than its theoretical maximum, it was more than doubled when compared to those of previously reported DNA-based data storage methods. Also, as mentioned in the previous chapter, the amount of NGS depth is several hundred times larger than previous works, when storing information of similar amount.

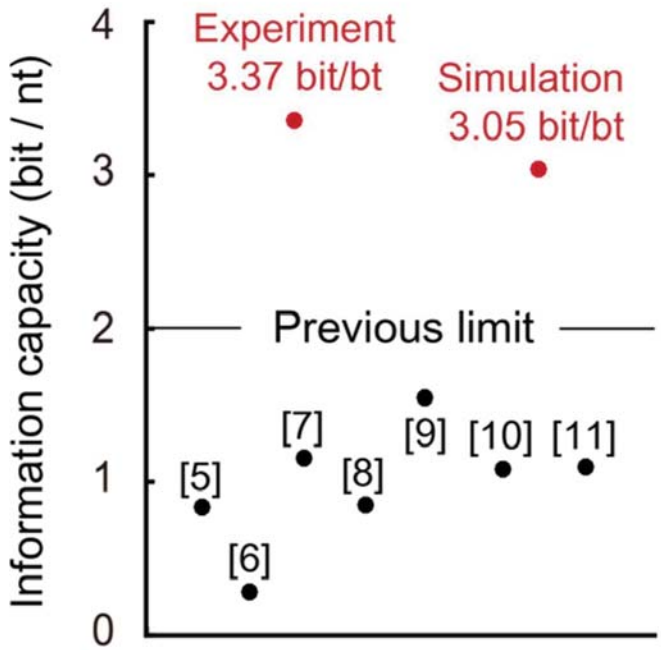


Figure 5.1 Information capacity achieved in this dissertation and comparision between capacity from previous researches. The dots in the graph describe the information capacity in previous research, and the numbers indicate the corresponding reference. This figure has been modified from the previous research[17].

	Grass *et al*.	Blawat et *al*.	Erlich *and* Zielinski	Organick *et al*.	This work (Experiment)	This work (Simulation)
Error correction method	RS	RS	Fountain	RS	RS	RS
Information capacity (bit/nt)	1.16	1.18	1.57	0.81	3.37	3.05
Physical information density (Pbytes/g)	25	-	214	-	772	-
Input data (Megabytes)	0.08	22	2.15	200.2	854byte	100
Number of oligonucleotides	4,991	1 million	72,000	13 million	45	-
Minimum NGS coverage (average)	372x	160x	10.5x	5x	250x	1300x

Table 5.1 Comparision between methods (Rep, Repetiotion method. RS, Reed-Solomon error correction): 10^6 bytes, Pbyte : Peta byte, 10^{15} bytes.

5.2. Cost Projection of the Platform

As shown in previous chapters, proposed method doubled information density when compared to the previous researches, although the proposed method requires higher NGS coverage. However, the sequencing technology has a rapid speed of development and the current DNA sequencing cost per base is approximately 50,000 times lower than the synthesis cost per base, when applying the synthesis cost based on the cost of column-based DNA synthesis reported in Erlich and Zielinski[10] and the DNA sequencing cost reported by K. Wetterstrand[24].

The cost of the storage is projected, when assuming the 1MB of the data is stored (Figure 5.2). The cost of synthesis can be determined by dividing 1MB by the information capacity and finding the number of nucleotides needed to be synthesized. For the projection, I used the information capacity of the simulation in Chpater 4.1, which can be used as large-scale data storage. First, since the information capacity does not reflect adapters on both ends of fragment, the value of 160/200 should be multiplied. The cost then can be calculated by number of nucleotide and the synthesis cost per nucleotide. Also, NGS cost could be estimated by multiplying the estimated number of nucleotides and the NGS coverage and its cost. When calculating NGS cost, 2000x depth is applied assuming extreme coverage. Finally, the cost of proposed platform is 2052 USD/1MB when using 15 encoding characters and

1795 USD/1MB when using 21 encoding characters, which is approximately half of the previous minimum of $3555/1MB. Also, NGS cost is less than 5% of synthesis cost. The price of DNA synthesis used here is a pool-based DNA synthesis, and since it is a column-based synthesis, minor modification of the synthesis machine will allow the use of degeneration base. From this, the expected price will be realized in real terms.

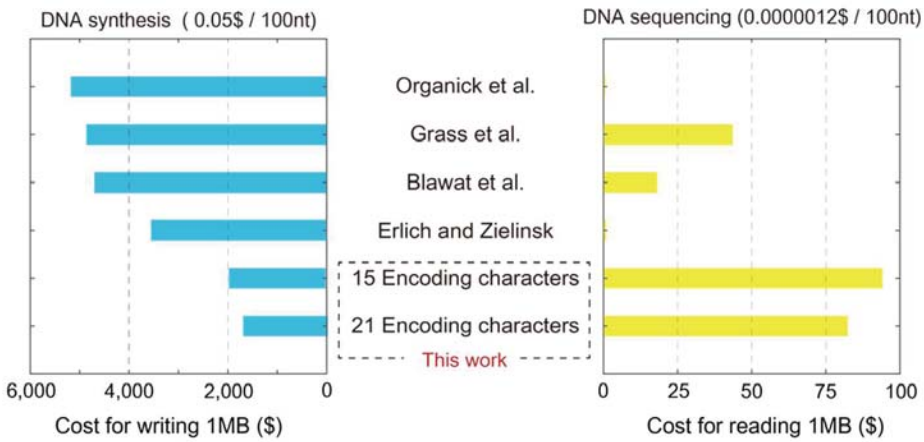


Figure 5.2 Cost comparision between DNA-based data storage methods: Of the proposed studies in the past, only describe the price of recovering data from DNA perfectly. The price was calculated by comparing the number of nucleotides synthesized and the amount of NGS from each study and multipying that with the cheapest synthesizer and analyst in the current market. For 15 encoding characters, A, C, G, T and all other eleven degenerate bases were used. Additionally, A, C, G, T, [R, Y, M, K, S, W – ratio of bases mixed of 3:7 and 7:3], H, V, D and N were used as 21 encoding characters. This figure has been modified from the previous research[17].

5.2.1. Outlook for Practical Use of DNA-based Data Storage

In this study, I proposed a method to reduce the price of DNA-based data storage by half. But now, the lowered price is still much higher than the cost of existing storage. For the method of increasing information capacity by introducing an additional degenerate base, the increase in the efficiency may not be significantly high because the information capacity increases as a log

function for the encoding character. If the ratio of the 11 degenerate bases is adjusted to increase the total number of characters to 32 and increased to 4 bit / nt, the DNA synthesis cost will be reduced to about 1000 $ and about 20% of this, 200% will be the NGS cost. From this point on, NGS prices will take up a large portion of the total cost, making it difficult to use additional degenerate bases. In this case, an additional reduction in synthetic and analytical prices of at least 100 times would be required, for practical use of the storage system. If the sequencing cost has been reduced by a factor of 100 for about two years and there is little decline in the synthesis cost(http://www.synthesis.cc/synthesis/2016/03/on_dna_and_transistors), a new synthetic technique may be needed.

5.3. Applicability of Degenerate Bases to Other DNA-based Data Storage Methods

In this dissertation, the degenerate base was only applied to DNA-codon-based encoding. However, the degenerate base could further be used in other data to DNA encoding method. In particular, in case of random matching of the DNA[5] and modified or similar version of DNA codon methods[6], [11], the method could be directly used. When the encoding method is changed, a dramatic increase of the information capacity is not expected because there is no dramatic difference in information capacity between each encoding method even when using the 4 bases.

For the case of the DNA fountain coding[10] which has the most efficient algorithm in terms of information capacity, since the code check the GC contents of the fragment, the degenerate that has the variance of the GC contents in single design could not be used directly. As an alternative, while using the fountain code, checking all GC variants of the single design, only passing the GC contents between 30~60% and checking the possibility of the decoding by using the computation could be the solution. For this, the efficient method for variant generation and filtering should be developed since there could be more than 10^{15} variants in the fragment longer than 100nt. Also, the computation power, time and its cost should also be added to the consideration of the method. Regard to this, research will be followed in near future.

5.4. Future Works

In this dissertation, by adding degenerate bases to the data to DNA encoding, the information capacity and physical density were more than doubled when compared to those of previously reported DNA-based data storage methods. To realize the projected cost, oligonucleotide pool synthesis setup that has synthesis capability for all degenerate sequence should be developed. Also, precisely control the ratio of the nucleotides in degenerate base, with a low deviation of nucleotide combinations could allow the higher information capacity.

Other than decreasing synthesis cost or increasing the information capacity, the length limitation and quality of DNA synthesis platforms also should be increased. Currently, there is a synthesis limit of about 200 nt, due to the low yield of the synthesis platform, which is synthesizing DNA by linking DNA blocks corresponding to one nucleotide one by one. If the length limit of the synthesis increases, the nucleotide corresponding to the address can be saved, thereby increasing the information capacity will be possible. Also, if the lowered error rate while synthesis is possible, data can be recovered even if the amount of error correction is reduced, which can increase the information capacity.

In addition to the cost problem that discussed in the dissertation, there are several problems to be solved in the near future. For example, the speed of data

reading and writing should also be accelerated through the development of platforms. Currently, in the case of 200 nt of DNA fragment, it takes about 2 days for data writing and a day for reading. Because all processes are done in high-parallel manner, current platforms are appropriate for processing large amounts of data at once, but they are not suitable for instant reading and writing of small amounts of information. For this, the cold data described in the Chapter 1 would be appropriate, before the problem is solved.

In addition, in order to develop DNA in a way that enhances convenience, it is necessary to overcome the material characteristics of DNA. For example, once a DNA is stored, certain information cannot be altered or deleted. To get a hint of a solution to this in nature, the CRISPR system[33] could be answer to the problem, since the system has the ability to erase and replace certain part of the gene. However, in this case, redundancy for current error correction is also required to be deleted or modified together, for this, all error correction codes previous introduced is not appropriate, since the redundancy is made from the combination of the fragments. Also, unlike conventional information storage methods that organize information in physical locations, DNA-based storage methods store information in random pieces as a powder form. In this case, it is difficult to find information physically. Even though the PCR can be used to amplify specific DNA, all data should be dissolved in the solution to recover only specific data, and the rest will be discarded. To solve this problem, each data should be physically controllable. The physically controllable DNA

system also could be found in the cells, that over-expresses the RNA for a particular bio-chemical input and preserves the original DNA-based gene.

In summary, if information is stored in a biosystem such as a cell, information can be provided according to a particular input, and the information can be modified or deleted. While these systems cannot be achieved with current technology, it is thought that this can be achieved by the development of synthetic biology and system biology field.

The following chapters describe approaches that can be addressed in the near future, on the few aspects of described problems.

5.4.1. Clustering of NGS Read for Shorter Fragment Decoding

In the NGS read filtering procedure, there is a large amount of data loss in the process of filtering an appropriate length of NGS read. This is due to deletion errors during DNA synthesis. Even though the length of DNA is shorter than designed, it still has the information, the required NGS coverage could be decreased if the shorter DNA is used for decoding. For this, clustering of NGS read (Figure 5.3)[11] that previous research used could be utilized. The method is rearranging the sequencing data by clustering and find the position of the deletion. However, since proposed platform in this thesis has high sequence heterogeneity variance in designed fragment, the modification of the clustering algorithm would be needed. Also, since the cost of NGS is negligible, the price drop by the clustering method will not be dramatic, the approach is not covered

in the thesis.

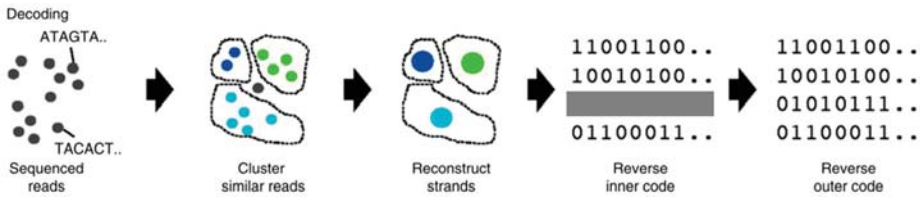


Figure 5.3 Clustering of sequencing reads for reconstructing the strands. By utilizing the method, shorter sequncing reads than design also could be used for data recovery. The figure has been modified from the previous research [11].

5.4.2. Addition of Inosine Base for DNA-based Data Storage

Inosine is the base that commonly found in tRNAs. Also, it could be synthesis using the current DNA synthesizer. The inosine is normally changing its base as G, while PCR based amplification, since the complementary of the base that polymerase recognize is the C. From this reason, the inosine cannot be used as the additional base.

However, the acrylonitrile treatment could read the chemical change of the base to not form the base pairing[34](Figure 5.4). So, by using high-fidelity enzyme, which could skip this base, the shorter DNA fragment could be generated by PCR amplification. From this, there is the length difference between fragment could be generated by the chemical treatment and it could be used as another alphabet for the DNA-based storage. Also, since the reduction of DNA length also could be confirmed in the gel electrophoresis process before

NGS or sequencing, Data could be instantly confirmed, and another application area could be opened after future development.

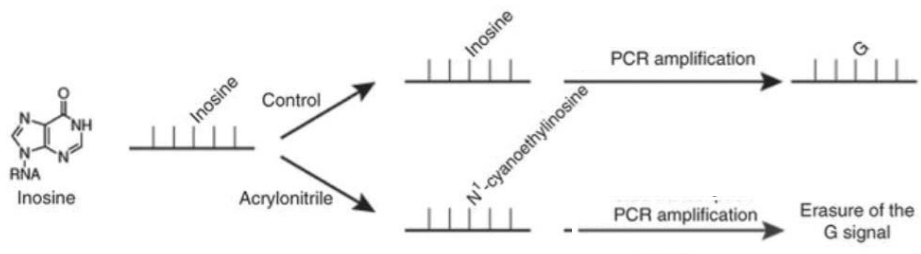


Figure 5.4 Chemical treatment to the inosine base could lead to length change of DNA fragment. The length difference could be used as another encoding alphabet for DNA-based data storage. The figure has been modified from the previous research [34].

5.4.3. Indexing of DNA on Encoded Microparticle

Even though the DNA-based data storage has its advantage in high physical information density, since its physical form is the water-soluble powder, the data could not be indexed in physical form. From the reason, one could not separate the specific. Also, the original copy of the DNA would be discarded after PCR amplification, the system is basically write once, read once (WORO) memory. Even though the random-access based data selection method, which utilize multiple primer set to enable PCR-based data selection, the original copy will be discarded and chance of reselection of the data would be

restricted.

In this chapter, I propose the microparticle-based DNA indexing method. By attaching DNA on the micro-sized encoded particle[35], DNA could be indexed(Figure 5.5). By using previous reported polymer based encoded particle, which also could carry the DNA by attaching, the original copy of the DNA could be reserved on the particle. The encoded patter, or the QR code could be give the index of the data and the adapter sequence of the DNA library. From this, the user could see the index by screening the QR codes and physically select the data. I selected the micro-sized polymer particle, since the hundreds of micrometers is the minimal size that could be seen and selected with bare eyes.

Also, since the original copy of the DNA is linked on the particle, there would not be the loss of the data, after the serial PCR-based amplification in solution. From this, selective file retrieve in multiple time could be done, by using the bead (Figure 5.6). Even if previous research claimed that the multiple read could be done by data amplification using PCR, the uneven representation of the data, or data bias will be accumulated and data error will be increased during the amplification, as see in Chapter 3. From this, proposed method could achieve the multiple read system that previous system could not, without bias (Figure 5.7).

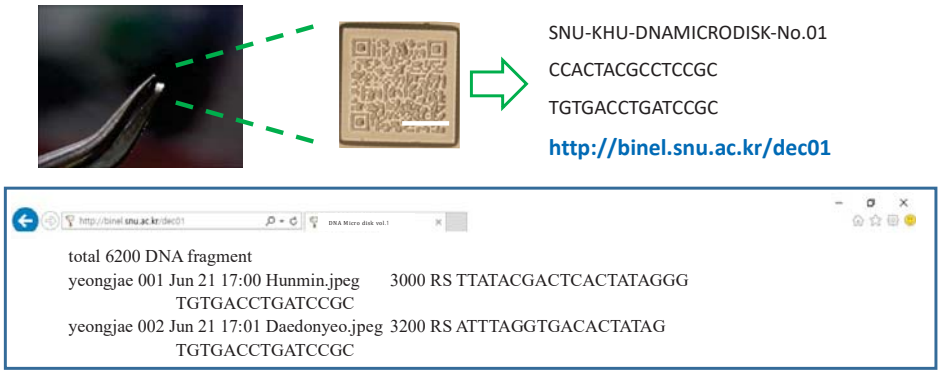


Figure 5.5 Digital data is encoded and synthesis to DNA and stored in the encoded microparticle. The QR code on the microparticle gives the brief information of the DNA library and the adapter sequence. Scale bar : 200um.

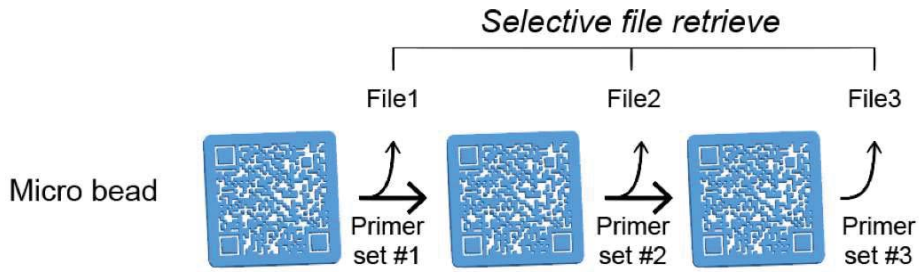


Figure 5.6 Selective file retrieve from proposed system.

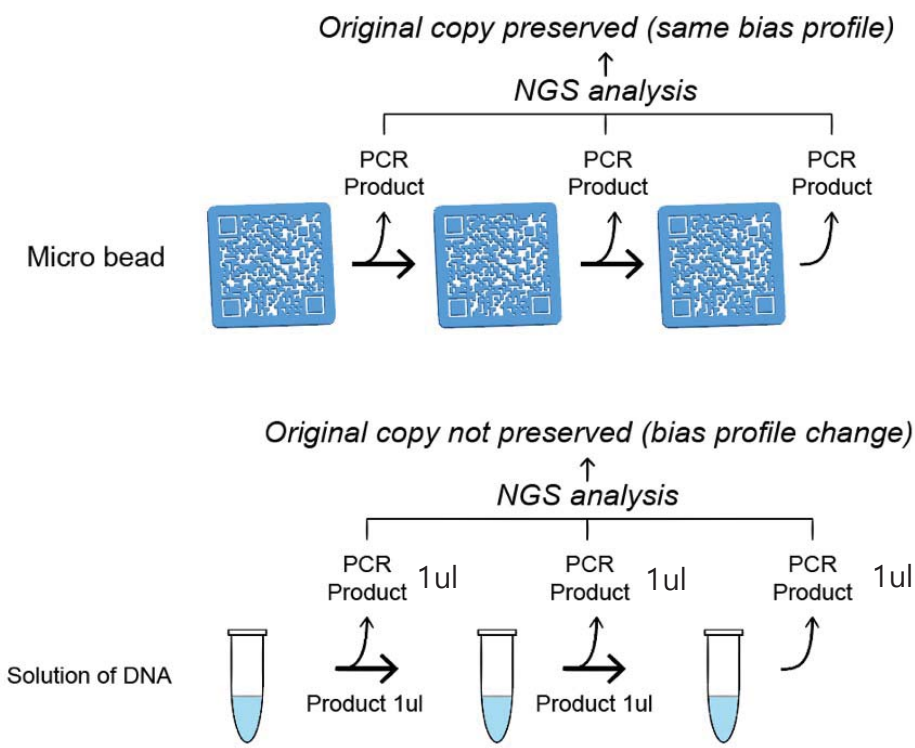


Figure 5.7 Multiple read of the data using the microparticle attached DNA, without the uneven representation of the data due to the PCR bias.

Chapter 6. Bibliography

[1] D. Reinsel, J. Gantz, and J. Rydning, “Data Age 2025: The Evolution of Data to Life-Critical Don’t Focus on Big Data; Focus on the Data That’s Big Sponsored by Seagate The Evolution of Data to Life-Critical Don’t Focus on Big Data; Focus on the Data That’s Big,” 2017.

[2] V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church, and W. L. Hughes, “Nucleic acid memory,” *Nature*, vol. 15, no. 4, pp. 366–370, Apr. 2016.

[3] Steve Campbell, “Navigating Storage in a Cloudy Environment,” *Cloud Expo 2013*, 2014. [Online]. Available: https://www.slideshare.net/hgststorage/navigating-storage-in-a-cloudy-environment. [Accessed: 07-May-2018].

[4] A. Mendoza, “Cold Storage in the Cloud : Trends , Challenges , and Solutions.”

[5] G. M. Church, Y. Gao, and S. Kosuri, “Next-generation digital information storage in DNA.,” *Science*, vol. 337, no. 6102, p. 1628, Sep. 2012.

[6] N. Goldman *et al.*, “Towards practical, high-capacity, low-maintenance information storage in synthesized DNA.,” *Nature*, vol. 494, no. 7435, pp. 77–80, Feb. 2013.

[7] 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.,” *Angew. Chem. Int. Ed. Engl.*, vol. 54, no. 8, pp. 2552–5, Feb. 2015.

[8] J. Bornholt *et al.*, “A DNA-Based Archival Storage System - Microsoft Research,” *ACM SIGOPS Operating Systems Review*,

25-Mar-2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2954680.2872397. [Accessed: 15-Feb-2016].

[9] M. Blawat *et al.*, "Forward Error Correction for DNA Data Storage," *Procedia Comput. Sci.*, vol. 80, pp. 1011–1022, 2016.

[10] Y. Erlich and Z. Dina, "DNA Fountain enables a robust and efficient storage architecture," *Science (80-.).*, no. 355, pp. 950–954, Mar. 2017.

[11] L. Organick *et al.*, "Random access in large-scale DNA data storage," *Nat. Biotechnol.*, Feb. 2018.

[12] H. Römpler *et al.*, "Nuclear gene indicates coat-color polymorphism in mammoths.," *Science*, vol. 313, no. 5783, p. 62, Jul. 2006.

[13] D. Paunescu, M. Puddu, J. O. B. Soellner, P. R. Stoessel, and R. N. Grass, "Reversible DNA encapsulation in silica to produce ROS-resistant and heat-resistant synthetic DNA 'fossils'.," *Nat. Protoc.*, vol. 8, no. 12, pp. 2440–8, Dec. 2013.

[14] A. Extance, "How DNA could store all the world's data," *Nature*, vol. 537, no. 7618, pp. 22–24, Aug. 2016.

[15] C. Bancroft, T. Bowler, B. Bloom, and C. T. Clelland, "Long-Term Storage of Information in DNA," *Science (80-.).*, vol. 293, no. 5536, p. 1763c–1765, Sep. 2001.

[16] Jay Greene, "Is DNA the Future of Data Storage?," *The Wall Street Journal*, 2016. [Online]. Available: https://www.wsj.com/articles/is-dna-the-future-of-data-storage-1477405351. [Accessed: 07-May-2018].

[17] Y. Choi *et al.*, "Addition of Degenerate Bases to DNA-based Data Storage for Increased Information Capacity," *bioRxiv*, p.

367052, Jul. 2018.

[18] S. L. Beaucage and R. P. Iyer, “Advances in the Synthesis of Oligonucleotides by the Phosphoramidite Approach,” *Tetrahedron*, vol. 48, no. 12, pp. 2223–2311, 1992.

[19] D. Aird *et al.*, “Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries,” *Genome Biol.*, vol. 12, no. 2, p. R18, Feb. 2011.

[20] M. G. Ross *et al.*, “Characterizing and measuring bias in sequence data,” *Genome Biol.*, vol. 14, no. 5, p. R51, May 2013.

[21] G. Ananda *et al.*, “Distinct Mutational Behaviors Differentiate Short Tandem Repeats from Microsatellites in the Human Genome,” *Genome Biol. Evol.*, vol. 5, no. 3, pp. 606–620, Mar. 2013.

[22] R. Williams, S. G. Peisajovich, O. J. Miller, S. Magdassi, D. S. Tawfik, and A. D. Griffiths, “Amplification of complex gene libraries by emulsion PCR,” *Nat. Methods*, vol. 3, no. 7, pp. 545–550, 2006.

[23] K. H. Hecker and R. L. Rill, “Error analysis of chemically synthesized polynucleotides.,” *Biotechniques*, vol. 24, no. 2, pp. 256–60, Feb. 1998.

[24] K. Wetterstrand, “DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP),” *Natl. Hum. Genome Res. Inst.*

[25] S. Kosuri and G. M. Church, “Large-scale de novo DNA synthesis: technologies and applications,” *Nat. Methods*, vol. 11, no. 5, pp. 499–507, Apr. 2014.

[26] Y. Zhang *et al.*, “A semi-synthetic organism that stores and retrieves increased genetic information,” *Nature*, vol. 551, no.

7682, pp. 644–647, 2017.

[27] C. Mayer, G. R. McInroy, P. Murat, P. Van Delft, and S. Balasubramanian, “An Epigenetics-Inspired DNA-Based Data Storage System,” *Angew. Chemie Int. Ed.*, Jul. 2016.

[28] A. Cornish-Bowden, “Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.,” *Nucleic Acids Res.*, vol. 13, no. 9, pp. 3021–30, May 1985.

[29] E. M. LeProust *et al.*, “Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process,” *Nucleic Acids Res.*, vol. 38, no. 8, pp. 2522–2540, May 2010.

[30] M. A. Quail *et al.*, “Optimal enzymes for amplifying sequencing libraries,” vol. 109, no. 1, 2012.

[31] S. O. Oyola *et al.*, “Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes.,” *BMC Genomics*, vol. 13, p. 1, Jan. 2012.

[32] J. Temsamani, M. Kubert, and S. Agrawal, “Sequence identity of the n-1 product of a synthetic oligonucleotide.,” *Nucleic Acids Res.*, vol. 23, no. 11, pp. 1841–4, Jun. 1995.

[33] R. Barrangou, “Cas9 Targeting and the CRISPR Revolution,” *Science (80-.).*, vol. 344, no. 6185, pp. 707–708, May 2014.

[34] C.-X. Song, C. Yi, and C. He, “Mapping recently identified nucleotide variants in the genome and transcriptome,” *Nat. Biotechnol.*, vol. 30, no. 11, pp. 1107–1116, Nov. 2012.

[35] S. Eun Chung *et al.*, “One-step pipetting and assembly of encoded chemical-laden microparticles for high-throughput multiplexed bioassays.,” *Nat. Commun.*, vol. 5, p. 3468, Jan. 2014.

Chapter 7. 국문 초록

DNA 기반 데이터 저장은 디지털 데이터를 DNA 서열로 변환한 후 합성하고 저장하는 기술이다. 이 기술은 DNA의 물리적 장점으로 인해 접근이 빈번하지 않은 백업 데이터 저장에 사용 될 수 있을 것으로 예측되고 있다. 본 기술의 장점은 다음과 같다. 첫째, 본 기술은 DNA를 이용하기에 수세기 동안 전원 공급이나 오류 보정이 필요 없이 유지되며 이는 데이터 보존을 위해 전원 공급이나 오류 보정이 필요한 기존 저장 매체와는 대조적이다. 둘째, DNA는 그램 당 수백 페타 바이트(petabyte, 10^{15} bytes) 를 저장할 수 있는 물리적 정보 밀도를 가지고 있으며, 이는 기존의 저장 방법보다 수천 배 더 높다. DNA 기반 데이터 저장에 대한 과거 연구의 주요 목표는 데이터 오류 또는 손실을 줄이기 위한 데이터 변환 알고리즘을 개선하는 것이었으며, 데이터-DNA 변환 및 오류 수정 기능에 대한 설계 규칙이 제안되었다.

본 분야의 연구 개발에 있어 필요되어지는 다음 단계는 데이터 저장 비용을 줄이고 실제 사용을 가능하게 하는 것이다. DNA 기반 데이터 저장 의 현재 비용은 1MB의 데이터를 저장 하는데 약 3500 달러이다. 이러한 높은 가격을 낮추기 위해, 본 학위 논문에서는

단위 DNA 당 저장할 수 있는 데이터의 양을 늘림으로써 DNA 기반 데이터 저장 비용을 줄이는 것이 제안 되었다. 본 논문에서 제안 된 개념은 디지털 정보를 변환하는데 사용되는 DNA 문자인 A, C, G, T 에 추가 변환 문자로 네 개 문자의 혼합 인 degenerate base를 사용하는 것이다. 이를 통해 $\log_2$ (인코딩 문자 수) 인 이론적 최대 정보 용량은 원래 4 개의 인코딩 문자에 11 개의 degenerate base 가 추가되어 $\log_2 4$에서 $\log_2 15$ (bit / nt)로 증가되었다. 즉, 데이터 저장에 필요한 DNA 길이는 4 문자 기반 시스템에 비해 절반 이상 감소되었다. 기존 연구는 알고리즘 최적화에 집중한 연구를 진행하였다면, 본 연구는 DNA 합성 공정을 이용한 새로운 접근법이라 할 수 있다.

제안 된 아이디어를 사용하여, Data to DNA 인코딩, 분자 생물학 기반 DNA 처리 및 데이터 디코딩에 이르는 DNA 기반 정보 저장의 전체 과정이 본 논문에서 시연 및 시뮬레이션 되었다. 또한 시뮬레이션 및 비용 예측에서 1MB를 저장하는 비용은 이전 비용에 비해 50 % 감소 할 것으로 예상되었다. 이는 데이터 저장에 필요한 DNA의 길이가 절반 이하로 줄어들어 DNA 합성 비용이 감소하기 때문이다.

본 학위논문에서 제안된 방법은 정보 변환 및 DNA 합성

단계의 간단한 수정 만 통한다면 기존에 제안 된 거의 모든 DNA 기반 데이터 저장 방법에 적용 될 수 있으며, 이를 통해 경제적 효율성을 높일 수 있다. 따라서 제안된 아이디어와 실험은 DNA 기반 저장의 실제 구현에 도움을 줄 수 있을 것으로 기대된다.

주요어: DNA 기반 정보저장, 정보 저장법, Degenerate Base, 분자생물학, 합성생물학

학번: 2012-23246