Syst Synth Biol (2008) 2:19–25
DOI 10.1007/s11693-008-9020-5

COMMENTARY

Stabilizing synthetic data in the DNA of living organisms

Nozomu Yachie · Yoshiaki Ohashi · Masaru Tomita

Received: 19 October 2008 / Revised: 19 November 2008 / Accepted: 25 November 2008 / Published online: 13 December 2008

Abstract Data-encoding synthetic DNA, inserted into the genome of a living organism, is thought to be more robust than the current media. Because the living genome is duplicated and copied into new generations, one of the merits of using DNA material is long-term data storage within heritable media. A disadvantage of this approach is that encoded data can be unexpectedly broken by mutation, deletion, and insertion of DNA, which occurs naturally during evolution and prolongation, or laboratory experiments. For this reason, several information theory-based approaches have been developed as an error check of broken DNA data in order to achieve data durability. These approaches cannot efficiently recover badly damaged data-encoding DNA. We recently developed a DNA data-storage approach based on the multiple sequence alignment method to achieve a high level of data durability. In this paper, we overview this technology and discuss strategies for optimal application of this approach.

N. Yachie · Y. Ohashi · M. Tomita
Institute for Advanced Biosciences, Keio University, Tsuruoka
997-0017, Japan

M. Tomita
e-mail: mt@sfc.keio.ac.jp

N. Yachie · M. Tomita
Systems Biology Program, Graduate School of Media and
Governance, Keio University, Fujisawa 252-8520, Japan
e-mail: nzm@sfc.keio.ac.jp

Y. Ohashi (✉) · M. Tomita
Biomedical Group, Human Metabolome Technologies, Inc.,
246-2 Mizukami, Kakuganji, Tsuruoka 997-0052, Yamagata,
Japan
e-mail: ohashi@humanmetabolome.com

Keywords DNA data storage · Sequence alignment · Polymerase chain reaction (PCR) · Error check · Error correction · Genetically modified organism (GMO)

> "Increasing performance of CPUs and memories will be squandered if not matched by a similar performance increase in I/O. While the capacity of Single Large Expensive Disks (SLED) has grown rapidly, the performance improvement of SLED has been modest. Redundant Arrays of Inexpensive Disks (RAID), based on the magnetic disk technology developed for personal computers, offers an attractive alternative to SLED, promising improvements of an order of magnitude in performance, reliability, power consumption, and scalability."
>
> David Patterson, Garth Gibson, and Randy Katz 1988 when they first proposed RAID[1]

Introduction

DNA consists of stable double-stranded long polymers of only four different nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). For living organisms, the primary role of DNA is long-term and inheritable data storage of genetic information, which is the set of blueprints to construct cellular components, such as RNA and protein molecules. The features of a DNA sequence are analogous to a digital data sequence. For this reason, recent studies have focused on the behavior of DNA in

[1] For quote, see: Patterson et al. (1988)

the application of artificial data memory. For example, by inserting a synthetic data-encoding DNA molecule into the genome of living organisms, the encoded data can be reproducible and inheritable in the small media vessel of the living cell. The data-storage media broadly used today, including paper, magnetic media, and silicon chips, are easily damaged by human-caused and spontaneously occurred accidents, and require constant time-consuming and laborious maintenance. The properties of DNA provide the potential for realization of long-term data storage, which can be maintained as archived data for hundreds to thousands of years (Bancroft et al. 2001; Cox 2001; Smith et al. 2003; Yachie et al. 2007; Wong et al. 2003). Since Clelland et al. (1999) demonstrated the encryption of hidden messages into DNA, the DNA-data encoding approach has essentially focused on DNA steganography, useful for the identification of genetically modified organisms (GMO) by its encoded signature or trademark in genomic DNA (Arita and Ohashi 2004; Heider and Barnekow 2007a, b; Heider et al. 2008; Wong et al. 2003). However, the artificial DNA sequences can be changed or degraded by laboratorial treatment during the preparation of synthetic data-encoded DNA, and by genetic mutations, deletions, and insertions in generations beyond those of the data-stored cells. Furthermore, there is the possibility of interfusion of sequencing errors when a DNA sequencer retrieves the encoded data. For these reasons, sophisticated methodology is required to ensure the robustness and durability of encoded data in order to realize the significant potential for DNA-mediated data storage for large amounts of important information over an extended period of time.

PCR-based readout and error checkable codes

Most of the data-storage methods based on DNA developed so far depend on the polymerase chain reaction (PCR) for encoding and readout of information. In the data-storage procedure of these methods, data sequence initially is converted into DNA sequence according to a rule, which is usually a set of encryption keys or 'codons' and is like the reverse procedure of the translation of RNA sequence into the amino acid sequence of protein. Two unique template DNA regions, corresponding to the forward and reverse primers for DNA amplification, then are added to the data-encoded region, one at each end (Fig. 1a). According to the sequence design of data-encoded DNA, double-stranded synthetic DNA is prepared and finally inserted into the junk region of genomic DNA. For readout of encoded data, within the mixture of the data-inserted genomic DNA and both primer sequences, the data-encoded DNA region is amplified by PCR and decoded by a DNA sequencer.

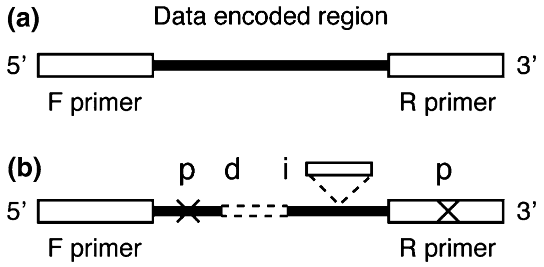

Fig. 1 Commonly used structure of data-stored synthetic DNA (**a**) and possible breakage points (**b**). 'F primer', forward primer region for DNA amplification; 'R primer', reverse primer region; 'p', point DNA breakage (mutation); 'd', deletion; 'i', insertion

The genomic DNA segment of the data-encoded region is altered, as diagrammatically represented in Fig. 1b, according to mechanisms underlying natural selective pressure in a living system. Moreover, in the data encoding and decoding procedures, data breakage can occur due to human error. Therefore, several approaches based on information theory have been proposed to check for errors in the DNA and broken DNA data (Arita 2004; Smith et al. 2003).

The comma code and the alternating code provide visible and regular rules for the nucleotide sequences of the data-encoded region. The comma code of DNA regularly repeats punctuation by using a single nucleotide within the data-encoded sequence and creates an automatic DNA reading frame (Smith et al. 2003). For example, a partial data-encoded DNA sequence utilizing the comma code is given by

$$G\_\_\_\_\_G\_\_\_\_\_G\_\_\_\_\_G, \tag{1}$$

where each G is the punctuation and A, C, and T are in the other spaces. Similarly, the redundant alternate code comprises a regularly alternating sequence of two divided groups of nucleotides, like

$$RYRYRY, \tag{2}$$

where R = A or G and Y = C or T. For both codes having regular rules of repetitive features in DNA sequences, errors are detectable only when the repetitiveness is ostensibly broken; the error check and correction is partially satisfied in each code.

A strategy employing a binary comma-free code for DNA memory has also been proposed; this code utilizes the binary comma-free and error-correctable data sequence in the DNA letters (Arita 2004). Although this comma-free code is robust and the error correction works to correct against small-scale damage such as DNA point mutations, this system does not have the capacity to recover broken data when a large DNA segment is deleted from the data-encoded DNA region.

A major disadvantage of these PCR-based methods is the susceptibility of the template DNA region for the introduction of errors. In the case when the forward or reverse primer regions are broken, there is the potential for the entire data sequence to become unreadable because of the failure to support PCR amplification. When this occurs, there is no code that can protect the template regions.

Concept of alignment-based readout

For more simple, flexible and durable data storage using DNA material, we recently proposed the alignment-based method (Yachie et al. 2007), which is independent of the PCR-based readout procedure but based on the sequence alignment method (Altschul et al. 1990). In this data-storage method, multiple oligonucleotide sequences encoding the same data are redundantly inserted into the genome of a living organism. When the encoded data is retrieved, the complete genome sequence, including multiple segments of inserted DNAs, is first read by a DNA sequencer, and then the multiple regions encoding the same data are searched for by sequence alignment. Based on the multiple alignments of sequences of data-encoded regions, errors that may have occurred within the encoded data can be checked and corrected.

Storing multiple-copied data

When multiple segments of the same DNA sequence exist within a single genome, or even within a single cell, there is the possibility for problems to occur with the storage of data. In particular, the existence of two of the same DNA sequences within a bacterial genome is known to induce homologous DNA recombination (Kowalczykowski et al. 1994; Kuzminov 1999). Homologous DNA recombination subsequently can disturb the growth and development of data-storage cells or the inserted DNA sequences may be removed from the genome in early cell division in cell culture. Therefore, in order to safely encode multiple-copied data sequences into genomic DNA, unique DNA oligomers encoding the same data according to different data encryption procedures are required.

Moreover, to realize the potential for this alignment-based method for storing and recovering multiple-copied data sequences, specific and universal behaviors are necessary for all the encryption transformations that generate unique DNA oligomers encoding the same data and for their corresponding decryption transformations. Here, we overview every set of reversible transformations of data encoding into and decoding from the DNA sequence. By the forward transformation function *f*, a given data sequence is encoded into the DNA sequence, and the encoded data can be completely read from the DNA sequence by the reverse transformation function f^{-1} (Fig. 2). Let L be the data sequence to be encoded, which is presented as a string of n letters $l_1 l_2 \ldots l_n$ $(n \geq 1)$. The data-encoded DNA sequence D, which is encoded from L as a string of ψ DNA letters $d_1 d_2 \ldots d_\psi$ $(\psi \geq 1)$, can be given by

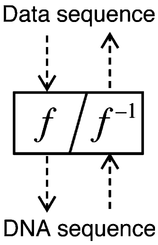

Fig. 2 Conceptual representation of reversible transformation for data sequence and DNA sequence. By using an encoding function, the data sequence is transformed into DNA sequence. Encoded data is decodable from the DNA sequence by f^{-1}

$$D = f(L). \tag{3}$$

Putting that the encoded data L in D is decodable by f^{-1}:

$$L = f^{-1}(D), \tag{4}$$

the reversibility of functions f and f^{-1} is represented by

$$f^{-1}(f(L)) = L. \tag{5}$$

Moreover, both f and f^{-1} satisfy the associative laws defined as

$$f(L_{i,j}) = f(L_{i,k})f(L_{k+1,j}), \tag{6}$$

$$f^{-1}(D_{p,q}) = f^{-1}(D_{p,r})f^{-1}(D_{r+1,q}), \tag{7}$$

where the partial segment from l_i to l_j in L can be represented as $L_{i,j} = l_i l_{i+1} \ldots l_j$ $(1 \leq i \leq j \leq n)$ and k satisfies $i \leq k < j$, and the partial segment from d_p to d_q in D can be represented as $D_{p,q} = d_p d_{p+1} \ldots d_q$ $(1 \leq p \leq q \leq \psi)$ and r satisfies $p \leq r < q$. Therefore, if the data-encoded DNA sequences have been partially altered, the sectional data segments can be decoded from the unchanged regions of DNA. We demonstrate an example procedure of sectional data read from a DNA sequence, which includes the partial segment of data-encoded DNA and two nonsense DNA sequences sandwiching it, as follows:

$$f^{-1}(\mathrm{N}\ldots\mathrm{N}f(L_{i,j})\mathrm{N}\ldots\mathrm{N}) = f^{-1}(\mathrm{N}\ldots\mathrm{N})L_{i,j}f^{-1}(\mathrm{N}\ldots\mathrm{N}), \tag{8}$$

where $f(L_{i,j})$ is the partial DNA region of data-encoded sequence derived from $L_{i,j}$ and N = A or C or G or T. When the DNA sequence $\mathrm{N}\ldots\mathrm{N}f(L_{i,j})\mathrm{N}\ldots\mathrm{N}$ is totally decoded by f^{-1}, the decoded sequence includes the sectional data $L_{i,j}$.

In order to store the data in the genome of a living organism, we used multiple different pairs of reversible transformation functions, with the behaviors described above. In the data-storage procedure, by using the multiple data-encoding functions, the data sequence is transformed to multiple different sequences of DNA (Fig. 3a). As

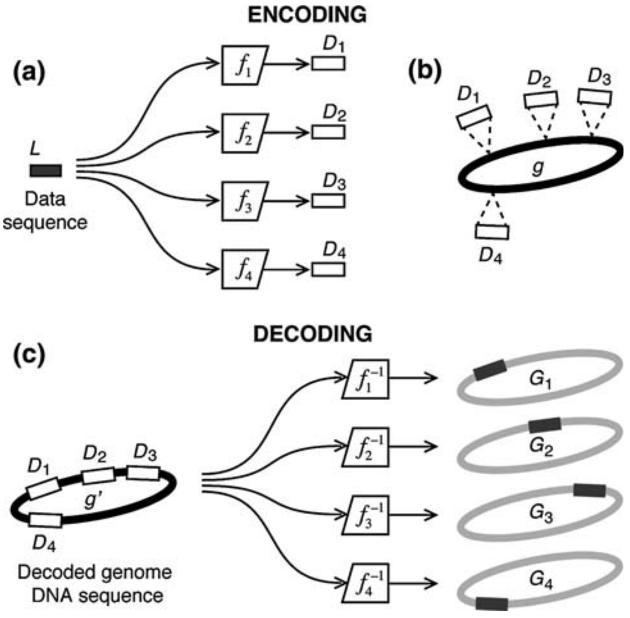

Fig. 3 Alignment-based DNA data storage and retrieval method. In this figure, the data storage and retrieval procedures are explained using four pairs of reversible encoding and decoding functions. When data L (dark gray box) are encoded into the genome of a living organism, four different DNA oligomers D_1, D_2, D_3 and D_4 are designed by querying the data sequence L to the encoding functions f_1, f_2, f_3 and f_4, respectively (**a**). Then, the double-stranded DNA oligomers are inserted into the genome g (**b**). In order to retrieve the encoded data L, the complete sequence of data-stored genome g' are initially decoded by a DNA sequencer, and the DNA sequence of g' is totally transformed to multiple data sequences G_1, G_2, G_3 and G_4 by the decoding functions f_1^{-1}, f_2^{-1}, f_3^{-1} and f_4^{-1}, respectively (**c**). According to our defined behaviors of the reversible functions, the same encoded data L must be included within each partial region of all the four decoded sequences (dark gray box). The encoded data L is thus retrievable by searching for the same data sequences by sequence alignment

described below, at least two different encodings is fundamental to this method, and at least three different encodings is optimal to ensure data durability. According to the design of the DNA sequence, double-stranded DNA oligomers that include the data-encoded sequences on one strand are synthesized and then inserted into the intracellular genomic DNA (Fig. 3b).

We previously demonstrated one of the simplest procedures to define the multiple and reversible transformations from a single data sequence to multiple DNA sequences (Yachie et al. 2007). In this procedure, a set of ‘codons’ is prepared that indicates the relationships between all possible patterns of x letters in a data sequence and their assigned DNA segments of the same size. There are x possible reading frames of ‘codons’ according to a one-by-one frame shifting of data letters in the target data sequence region, thus x different DNA sequences can be designed from the data sequence. This process mimics the DNA codons used for intracellular protein synthesis. There are three possible reading frames of three-letter DNA codons in the DNA sequence to encode the amino acid sequence of the protein.

Data retrieval by sequence alignment

In the data retrieval procedure, the complete genomic sequence harboring multiple synthetic DNA oligomers is fully sequenced by using a DNA sequencer, and then, the total sequence of genomic DNA is decompressed to multiple data sequences by using the decoding functions that are paired with the respective encoding functions used for data storage (Fig. 3c). The majority of regions of the respective long sequences decoded at the genomic level are nonsense, and they are mostly different from each other, because the different decoding functions are performed for a single genomic sequence. According to Eq. 8, the data sequence encoded in each synthetic DNA region of the genome accurately appears within the partial region of certain long data sequence transformed from the genomic sequence by the decoding function, which is the reverse of the encoding function used for the design of the respective region. Therefore, if all the data-encoded regions are not broken by DNA errors, every long decoded sequence must include the same unique data sequence in its partial region (Fig. 3c). By progressing through the series of data handling procedures, it is possible to search for and finally read out the same data sequence of encoded data by using the sequence alignment function.

Error check and correction by sequence alignment

At the end of the readout procedure, data durability can be further enhanced by taking advantage of the sequence alignment method. Because of the associative rules in the encoding and decoding functions defined in Eqs. 6 and 7, DNA mutations, deletions, and insertions of synthetic DNA are the causal factors of point breakage of data sequence, sectional data deletion, and nonsense data insertion, respectively. The types and positions of DNA errors are directly related to the errors in the decompressed data sequence. Therefore, according to this rule, even if some DNA errors are randomly contained in the multiple synthetic regions of data-stored genomic DNA, we can find the multiple-copied but partially broken data sequences by searching for similar data sequences in the respective long data sequences decoded from the genome, and the mismatches of aligned data sequences can identify the position of broken data-encoded sequences. Accordingly, the multiple-copied data sequences encoded within the different features of DNA sequences can fulfill an error check function.

Moreover, when more than three synthetic DNA oligomers are used for data storage, there is a high potential for correction of the identified data breakage points (Fig. 4). The natural DNA error rate in the genome of a living organism or in laboratorial experiments is not as high as the error rate associated with the insertion of artificial DNA sequences. Thus, it is extremely rare for the occurrence of both sources of errors at the same position within the multiple data-encoded regions. When the letters at a same position of multiple alignments of decoded data sequences are different, it is likely that the minority letters have arisen from data breakages. In this case, the broken data can be altered according to the majority decision rule. Consequently, for sequence alignment-based DNA data storage, data breakages can be identified and corrected by comparison with the same positions of other aligned sequences of decoded data.

When the number of decompressed data sequences from the synthetic DNA regions is two, the majority decision rule cannot be justified, and only the error check can be performed. At least three synthetic DNAs are necessary for error correction of broken data, and the cost of higher redundancy must be paid to achieve more robust durability of stored data, leading to a trade-off relationship between data durability and the copy number of encoded data. More durable data storage with stronger error correction can be achieved by the insertion of more synthetic DNA into the genome. This trade-off appears to favor data durability. By performing a computational simulation of data retrieval using this alignment-based method, we previously suggested that when the data storage is conducted by four synthetic DNAs and when as much as 15% of the data-stored genome is randomly mutated or deleted, the recovery rate is over 99% for data rescue (Yachie et al. 2007). According to combinatorial theory-based logic, the multiplication of synthetic DNA can improve data durability at an exponential rate.

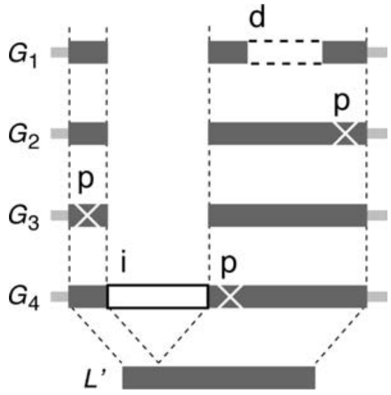

Fig. 4 Error correction of decoded sequences by sequence alignment. Encoded data surrounding regions of G_1, G_2, G_3 and G_4 (for detail, see footnote to Fig. 3) are represented with examples of their possible data breakage points. Dark gray boxes indicate data-encoding regions. ‘p’, point data breakage (or small breakage); ‘d’, sectional data deletion; ‘i’, sectional data insertion. According to the multiple alignment of decoded data sequences, each data breakage point can be identified and corrected by comparison with other multiple sequences, and the recovered data sequence L' can be retrieved

Cost and practical realization

For data storage into magnetic media, especially hard disks, redundant arrays of inexpensive disks (RAID) are used frequently (Patterson et al. 1988). The concept of RAID is to promote redundantly encoded data within a larger volume of inexpensive memory storage in order to secure data durability. However, when RAID technology was firstly proposed in 1988, the cost of memory storage was high when compared with the current technology, and the RAID concept of ‘inexpensive disks’ was considered novel.

The DNA data-storage methods prior to RAID in the computational technology field with PCR-based data retrieval process were associated with restraints including the high cost of DNA oligomer synthesis and DNA sequencing and the size limitation of artificial DNA that can insert into the genome of a living organism. The PCR-based methods store nonredundant data into one short synthetic DNA fragment along with the template DNA regions at each end. The cost of sequencing the partial data-encoded genomic region amplified by PCR is markedly lower than the cost of complete genome sequencing. By comparison, the alignment-based method requires the redundant encoding of data into multiple synthetic DNA molecules and sequencing at the genomic level. However, DNA sequencing is becoming less time-consuming and the associated cost of DNA synthesis and sequencing is decreasing.

There are several hundred complete sequences of chromosomal DNA from eukaryotes, prokaryotes, and archaea currently available via the public databases (Liolios et al. 2008). This area of research has fuelled a growing demand for higher speeds of DNA sequencing and lower cost and has boosted the emergence of new technologies (Church 2006; Hall 2007). For example, the recent development of sequencing technology utilizing emulsion PCR (Margulies et al. 2005; Shendure et al. 2005) has dramatically increased the speed of sequencing. Other new technologies are in development, and in the near future, new machines with the capacity to read one million bases per second at a low cost will be available (Bonetta 2006).

In the alignment-based method, following the design of multiple different DNA sequences encoding the same data, we can add the template DNA regions to each end of respective sequences and insert these into the living genome. Complete genome sequencing is unnecessary with this approach, because the partial genomic regions of respective synthetic DNA can be amplified by PCR and sequenced, and the multiple alignments of these sequences can check and correct errors. However, as explained above, PCR-based data retrieval is associated with the disadvantage of introducing breakage points in the template DNA region, and the readout procedure depends on appropriate or error-proof PCR amplification. Therefore, although data storage durability can be improved to a degree by multiple PCR amplifications and further by alignment of data sequences only from the successfully amplified DNAs, our proposed method fully utilizes the ability of the combinatorial and compositive error supplements by the multiple alignment of all the multiple-copied data, thus the data durability achieved by sequencing at the whole genomic level is beyond comparison. One of better strategies is that, with the combination of PCR-based readout, the complete genome sequencing is performed when the multiple readouts by PCR are failed. Optimal data durability can be guaranteed only by complete genome sequencing independent of PCR-based readout.

Although the length of synthetic DNA sequences that can be inserted into the genome of a living organism has been considered to be limited (Cox 2001), our method expends multiple and redundant synthetic DNA oligomers to copy-and-paste the data for storage. The upper limit of the total size of synthetic DNAs that can be inserted into the genome of a living organism has been increased remarkably by the development of megacloning technologies demonstrated in the transportation of whole bacterial genomes (Itaya et al. 2005, 2008). Notably, the entire 16.3-Kb mouse mitochondrial genome (Itaya et al. 2008), the 134.5-Kb rice chloroplast genome (Itaya et al. 2008), and the 3.5-Mb cyanobacterium genome (Itaya et al. 2005) have been inserted into the 4.2-Mb genome of *Bacillus subtilis* employed as a mother vessel. *B. subtilis* has the ability to form a tough and protective endospore, which allows this organism to survive extreme environmental conditions. Surprisingly, previous studies isolated strains of *Bacillus* species from an extinct bee trapped in 25–30 million-year-old amber (Cano and Borucki 1995) and from a brine inclusion within a 250 million-year-old salt crystal of the Permian Period (Vreeland et al. 2000). The development of megacloning technology utilizing *B. subtilis* has provided the potential for large-size and long-term DNA data storage.

In this paper, we overviewed the application of reversible transformations for data encoding and decoding for DNA data storage based on sequence alignment and complete genome sequencing. Because the data-encoding procedure for DNA requires no code and only the application of the associative rule, other codes, including the comma, alternate, and comma-free codes, can be used with our method. For example, the Huffman code is an economical code, in which is based on the varied frequencies at which English letters are used in the English-speaking sphere. The most frequent letter is ‘e’, with a frequency of 12.7%, and the least frequent letters are ‘q’ and ‘z’, both of which have a use frequency of 0.1% (Smith et al. 2003). According to this information, the Huffman code assigns a shorter ‘codon’ of nucleotide bases for more frequently used characters, and vice versa. Sophisticated codon design is necessary for the Huffman code: once the decoding procedure has commenced from the beginning of data-encoded DNA, there must be only one way to decompress the letters from the sequence of ‘codons’. For example, if the shortest DNA codon T is defined for the English letter ‘e’, no other longer codon can start with T (Smith et al. 2003). Similarly, in the alignment-based method, the encoded data within the synthetic DNAs can be short and durable, because of the versatile ability to be combined with previously proposed codes.

Conclusion

Similar to the general concept for magnetic disk drives, the alignment-based DNA data-storage method employs the redundant copy-and-paste-and-paste concept for storing data to realize the long-term DNA data storage of large amounts of important information in the small media vessel of the living cell. Although this methodology requires the use of redundant synthetic DNAs encoding the same data and whole-genome sequencing of the data-stored cell at high cost, research demands are promoting the development of new megacloning technology and high-speed DNA sequencing at a lower cost. For this reason, our proposed simple and flexible strategy may offer a practical solution for highly durable data storage in DNA.

Acknowledgments We are grateful to Kazuhide Sekiyama and Junichi Sugahara for helpful discussions and Dr. Kazuharu Arakawa for useful suggestions in the preparation of the manuscript. This study was supported by research funds from Yamagata Prefecture and Tsuruoka City to Keio University and a grant from the Japan Society for the Promotion of Science (JSPS) to N. Y.

References

Altschul SF et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Arita M (2004) Comma-free design for DNA words. Commun ACM 47:99–100. doi:10.1145/986213.986244

Arita M, Ohashi Y (2004) Secret signatures inside genomic DNA. Biotechnol Prog 20:1605–1607. doi:10.1021/bp049917i

Bancroft C, Bowler T, Bloom B, Clelland CT (2001) Long-term storage of information in DNA. Science 293:1763–1765. doi:10.1126/science.293.5536.1763c

Bonetta L (2006) Genome sequencing in the fast lane. Nat Methods 3:141–147. doi:10.1038/nmeth0206-141

Cano RJ, Borucki MK (1995) Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. Science 268:1060–1064. doi:10.1126/science.7538699

Church GM (2006) Genomes for all. Sci Am 294:46–54

Clelland CT, Risca V, Bancroft C (1999) Hiding messages in DNA microdots. Nature 399:533–534. doi:10.1038/21092

Cox JP (2001) Long-term data storage in DNA. Trends Biotechnol 19:247–250. doi:10.1016/S0167-7799(01)01671-7

Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol 210:1518–1525. doi:10.1242/jeb.001370

Heider D, Barnekow A (2007a) DNA-based watermarks using the DNA-Crypt algorithm. BMC Bioinformatics 8:176. doi:10.1186/1471-2105-8-176

Heider D, Barnekow A (2007b) DNA watermarks: a proof of concept. BMC Mol Biol 9:40. doi:10.1186/1471-2199-9-40

Heider D, Kessler D, Barnekow A (2008) Watermarking sexually reproducing diploid organisms. Bioinformatics 24:1961–1962. doi:10.1093/bioinformatics/btn342

Itaya M, Tsuge K, Koizumi M, Fujita K (2005) Combining two genomes in one cell: stable cloning of the *Synechocystis* PCC6803 genome in the *Bacillus subtilis* 168 genome. Proc Natl Acad Sci USA 102:15971–15976. doi:10.1073/pnas.0503868102

Itaya M, Fujita K, Kuroki A, Tsuge K (2008) Bottom-up genome assembly using the *Bacillus subtilis* genome vector. Nat Methods 5:41–43. doi:10.1038/nmeth1143

Kowalczykowski SC et al (1994) Biochemistry of homologous recombination in *Escherichia coli*. Microbiol Rev 58:401–465

Kuzminov A (1999) Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage lambda. Microbiol Mol Biol Rev 63:751–813

Liolios K et al (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 36:D475–D479. doi:10.1093/nar/gkm884

Margulies M et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Patterson D, Gibson G, Katz R (1988) A case for redundant arrays of inexpensive disks (RAID). Proc 1988 ACM SIGMOD Conf, vol 1. pp 109–116

Shendure J et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732. doi:10.1126/science.1117389

Smith GC, Fiddes CC, Hawkins JP, Cox JP (2003) Some possible codes for encrypting data in DNA. Biotechnol Lett 25:1125–1130. doi:10.1023/A:1024539608706

Vreeland RH, Rosenzweig WD, Powers DW (2000) Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. Nature 407:897–900. doi:10.1038/35038060

Wong PC, Wong K, Foote H (2003) Organic data memory using the DNA approach. Commun ACM 46:95–98. doi:10.1145/602421.602426

Yachie N et al (2007) Alignment-based approach for durable data storage into living organisms. Biotechnol Prog 23:501–505. doi:10.1021/bp060261y