NOTES

Alignment-Based Approach for Durable Data Storage into Living Organisms

Nozomu Yachie,[†,‡] Kazuhide Sekiyama,[†,‡] Junichi Sugahara,[†,§] Yoshiaki Ohashi,*[,†,||] and Masaru Tomita[†,‡,§,||]

Institute for Advanced Biosciences, Keio University, 403-1 Nipponkoku, Daihoji, Tsuruoka, 997-0017 Yamagata, Japan; Bioinformatics Program, Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa, 252-8520 Kanagawa, Japan; Department of Environmental Information, Keio University, 5322 Endo, Fujisawa, 252-8520 Kanagawa, Japan; and Human Metabolome Technologies, Inc., 246-2 Mizukami, Kakuganji, Tsuruoka, 997-0052 Yamagata, Japan

The practical realization of DNA data storage is a major scientific goal. Here we introduce a simple, flexible, and robust data storage and retrieval method based on sequence alignment of the genomic DNA of living organisms. Duplicated data encoded by different oligonucleotide sequences was inserted redundantly into multiple loci of the *Bacillus subtilis* genome. Multiple alignment of the bit data sequences decoded by *B. subtilis* genome sequences enabled the retrieval of stable and compact data without the need for template DNA, parity checks, or error-correcting algorithms. Combined with the computational simulation of data retrieval from mutated message DNA, a practical use of this alignment-based method is discussed.

Introduction

DNA has often been proposed to be a reproducible, heritable media in which a substantial amount of information can be included in its nucleotide sequence, while paper, magnetic media, and silicon chips are used for data storage today. Since all of these media are easily destroyed by personal and natural accidents, they are in need of endless attention to maintain their contents. Development of a DNA memory technology utilizing living organisms is of much greater potential than any of the existing counterparts to render a service of inheriting data. The robustness of DNA data, however, ensures the maintenance of archived information over extensive periods of time (hundreds to thousands of years) (*1*−*4*).

Message DNA has been used in steganography (*5*), as a means of short trademarks/signatures (*6*) and long-term storage (*3*, *7*). For economical and secure data encoding, sophisticated designs of DNA message regions and template primer sequences based on the polymerase chain reaction (PCR) method have been reported for data retrieval (*4*, *8*). Various codes have been developed for encryption in DNA sequences, including the economical Huffman code based on Huffman's algorithm that is used for short-term data storage (*4*). Similar to the parity-bit operation within barcodes of goods, the comma code uses a single base to punctuate the message and create an automatic DNA reading frame, while the redundant alternate code comprises an alternating sequence of purines and pyrimidines; these codes utilize DNA durability to archive long-term data (*4*). Retrieval of these forms of data requires PCR-based amplification, using template primer sequences, and DNA sequencing. In order to avoid misreading caused by wobble DNA annealing, a comma-free DNA design with attention to template DNA was proposed (*8*). However, these codes are not fully essential for the encoded DNA region to offer some safekeeping against mutation, which might make alternations to the sense of encoded data (*4*). Degradation of data could occur during DNA preparation in the laboratory, during storage, or at data retrieval (*4*). Additionally, genetic evolution and adaptation of the host organism could lead to the accumulation of mutations such as point mutations, insertions, and deletions in chromosomal DNA, resulting in destruction of data inheritance.

Magnetic disk drives store data securely by duplicating redundant data or parity codes into different sectors; data are then retrieved by the assembly of partial sectors. Here, we propose a method to copy and paste data within the genomic sequence of a living organism, *Bacillus subtilis*, thus acquiring versatile data storage and the robustness of data inheritance. Two or more different features of oligonucleotide DNA molecules compressed from the same data with an economical code were duplicated into multiple genomic loci for data storage. The encoded data is then retrievable by complete genome sequencing and searching for duplicated coding regions using multiple alignments of all the possible decoded bit-data sequences of the genomic DNA. Data durability was further discussed and estimated by computational simulation. We propose that this alignment-based method will be most beneficial in utilizing DNA as trademarks/signatures of living modified organisms (LMOs) and as valuable heritable media.

Materials and Methods

Conversion of Messages into Genetic Sequences. We encoded the message "E=mc^2 1905!" into the *B. subtilis* genome. The message was initially converted into binary

* To whom correspondence should be addressed. Tel: +81-235-29-0526. Fax: +81-235-29-0529. E-mail: ohashi@sfc.keio.ac.jp.
† Institute for Advanced Biosciences, Keio University.
‡ Graduate School of Media and Governance, Keio University.
§ Department of Environmental Information, Keio University.
|| Human Metabolome Technologies, Inc.

10.1021/bp060261y CCC: $37.00
Published on Web 01/25/2007

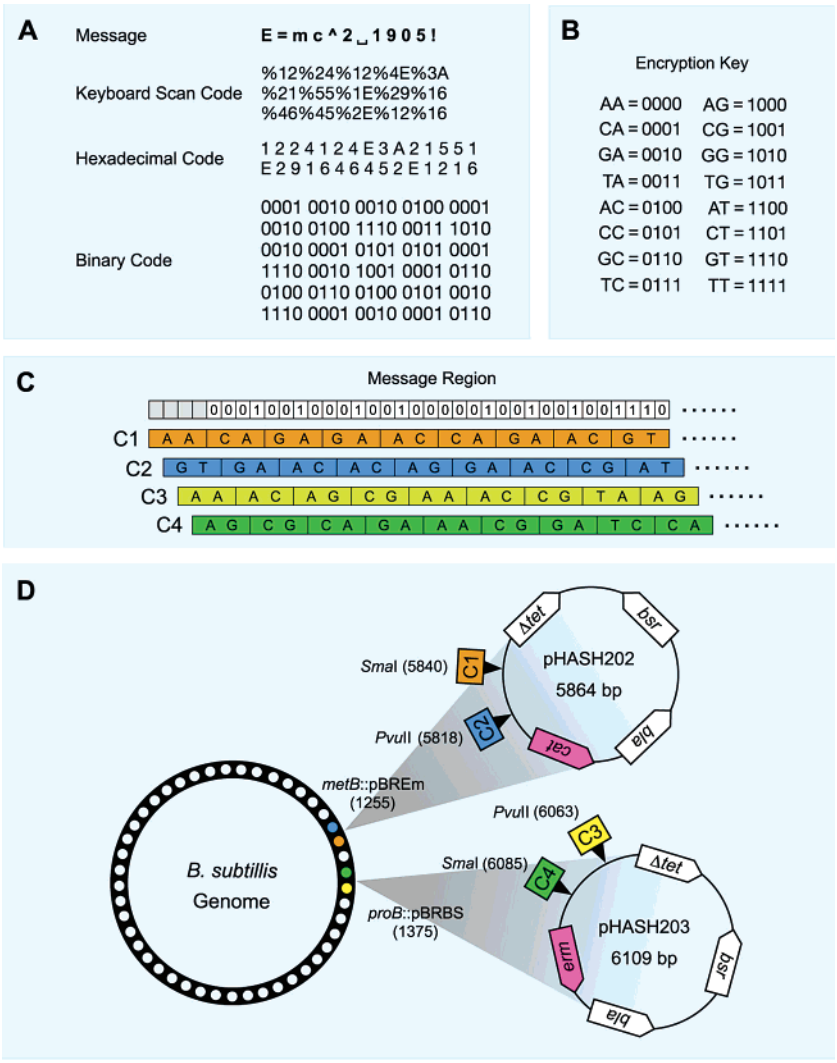

Figure 1. Encryption of a message in the *B. subtilis* genome. (A) The message "E=mc^2 1905!", including the "shift" and "space" keys, was converted into hexadecimal code (separated by %) according to the Keyboard Scan Code (Set2). The binary code was then produced by four-bit representation of each hexadecimal number. (B) The encryption key used to encode the message in DNA. Each pattern of four-bit binary code is represented by dinucleotides. (C) DNA cassettes C1, C2, C3, and C4 were translated by one-bit by one-bit frame shifting. Two sets of dinucleotides at the 5′ and 3′ regions encode different binary codes (see Table 1 for details). (D) Cassettes were subcloned into pHASH202 or pHASH203 plasmids and inserted into the *B. subtilis* genome.

sequence according to the make signals of the Keyboard Scan Code Set2 (*9*) that represent all keyboard inputs (Figure 1A). The hexadecimal codes generated from keyboard inputs were shifted into bit data, and the encryption keys translated each four-bit binary code into dinucleotides (Figure 1B). Because every encryption key has the reading frame size of four-bit, it is possible to translate a bit data sequence into four multiple oligonucleotide sequences in the four different reading frames (Figure 1C). In this study, four of all of the possible different oligonucleotide cassettes, C1, C2, C3 and C4, were designed for the data storage of the message "E=mc^2 1905!" (Table 1).

Introduction of Data into Plasmid Vectors. Introduction of the message sequences, C1 to C4, was carried out as summarized in Figure 1D. The chemically synthesized oligonucleotides C1 and C2 were annealed and subcloned into the T-extended *Sma*I and *Pvu*II sites, respectively, of pHASH202 (pSPIBER01); C3 and C4 were cloned into the *Sma*I and *Pvu*II sites, respectively, of pHASH203 (pSPIBER02) (*10*). *Escherichia coli* strain DH5α was used as the cloning host. The sequence of each DNA cassette was confirmed using the DNA sequencer ABI3100 (Applied Bioscience, CA).

Transformation of *B. subtilis.* *B. subtilis* competent cells were prepared and transformed by the two-step culture method (*11*). In order to introduce the cassettes, *B. subtilis* BEST2136 was used as a recipient (Table 2). This strain is resistant to erythromycin (Em), blasticidin S (BS), and tetracycline (Tc). The GB01 cells were generated by a mixture of 0.1 μg mL^{-1} of BEST2136 genomic DNA and 1 μg mL^{-1} of the pSPIBER01 plasmid. The chloramphenicol (Cm)-resistant transformants were selected on LB medium plates containing 5 μg mL^{-1} of Cm. The colonies that appeared after incubation at 37 °C overnight displayed an Em-sensitive and BS/Cm/Tc-resistant phenotype, indicating that the cassette C1 and C2 had been introduced within the *metB*::pBREm locus in the genome. Then, the GB02 cells were generated by a mixture of 0.1 μg mL^{-1} of GB01 genomic DNA and 1 μg mL^{-1} of the pSPIBER02 plasmid. The Em-resistant transformants were selected on LB medium plates containing 5 μg mL^{-1} of Em. After overnight incubation at 37 °C, the cells were BS-sensitive and Cm/Em/Tc-resistant, indicating that the cassettes C3 and C4 had been introduced within the *proB*::pBRBS locus in the genome. The four sequences of cassettes within the genomic DNA of the data-

Table 1. Oligonucleotide Sequences Encoding "E=mc^2 1905!"

cassette	oligonucleotide sequence[a]	encoded bit data
C1	AA,CA,GA,GA,AC,CA,GA,AC,GT,TA,GG,GA, CA,CC,CC,CA,GT,GA,CG,CA,GC,AC,GC,AC, CC,GA,GT,CA,GA,CA,GC,TT (64 nt)	0000,0001,0010,0010,0100,0001,0010,0100,1110,0011,1010,0010,0001, 0101,0101,0001,1110,0010,1001,0001,0110,0100,0110,0100,0101,0010, 1110,0001,0010,0001,0110,1111 (128 bit)
C2	GT,GA,AC,AC,AG,GA,AC,CG,AT,TC,AC,AC, GA,GG,GG,TA,AT,CC,GA,GA,AT,AG,AT,AG, GG,CC,AT,GA,AC,GA,AT (62 nt)	1110,0010,0100,0100,1000,0010,0100,1001,1100,0111,0100,0100,0010, 1010,1010,0011,1100,0101,0010,0010,1100,1000,1100,1000,1010,0101, 1100,0010,0100,0010,1100 (124 bit)
C3	AA,AC,AG,CG,AA,AC,CG,TA,AG,GT,AG,AG, CC,CC,AC,TC,AG,GG,AC,CC,CG,CA,CG,CA, AC,TG,AG,AC,AG,CC,AG (62 nt)	0000,0100,1000,1001,0000,0100,1001,0011,1000,1110,1000,1000,0101, 0101,0100,0111,1000,1010,0100,0101,1001,0001,1001,0001,0100,1011, 1000,0100,1000,0101,1000 (124 bit)
C4	AG,CG,CA,GA,AA,CG,GA,TC,CA,CT,CA,AA, GG,GG,AG,TT,CA,AC,AG,TG,GA,TA,GA,GA, CG,TC,AA,CG,AA,TG,TC (62 nt)	1000,1001,0001,0010,0000,1001,0010,0111,0001,1101,0001,0000,1010, 1010,1000,1111,0001,0100,1000,1011,0010,0011,0010,0010,1001,0111, 0000,1001,0000,1011,0111 (124 bit)

[a] Nucleotide sequences are given in the 5′ → 3′ direction. Dinucleotides separated by commas indicate encryption keys transformed by 4-bit binary codes. The 5′ and 3′ outer regions of bit data sequences (underlined) were designed differently to identify initiations and terminations of encoded data by sequence alignment.

Table 2. Bacterial Strains Used

strain	relevant genotype	description[a]
BEST2136	*metB*::pBREm *proB*::pBRBS	recipient strain, Em/BS/Tc-resistant
GB01	*metB*::*cat*::(C1, *C*2) *proB*::pBRBS	Em-sensitive, Cm/BS/Tc-resistant, with cassette C1 and C2
GB02	*metB*::*cat*::(C1, *C*2) *proB*::*erm*::(C3, C4)	BS-sensitive, Em/Cm/Tc-resistant, with cassette C1, C2, C3, and C4

[a] BS, blasticidin S; Cm, chloramphenicol; Em, erythromycin; Tc, tetracycline.

encoded strain GB02 were confirmed by DNA sequencing (ABI3100, Applied Bioscience, CA).

Estimation of Data Durability by Computational Simulation. In order to estimate the durability of the data obtained with our method, we computationally simulated data retrievals and calculated the data recovery rates from mutated and/or deleted cassette sequences. We prepared 10,000 sets including four pairs of different nucleotide cassettes and their corresponding bit data sequences harboring the same data region, varying in length from 100 to 10,000 bits. The nucleotide sequences were randomly mutated and deleted of their partial fragments in silico, and their corresponding bit data sequences were changed. Then, in each set, the mutation/deletion rate of the total of the four cassette sequences was calculated, and the data recovery rate was estimated; we calculated the percent of positions where less than half of the bit characters were changed in the data region. Conducting the same procedure, the recovery rates of the data encoded by two and three cassettes were calculated.

Results and Discussion

The concept presented here is the use of sequence alignment to retrieve encoded data that are duplicated in separate genomic loci. For data storage, two or more sequences with different oligonucleotide features are compressed and copied from the same bit data through multiple data-to-nucleotide translations and then inserted into the genomic DNA. Data retrieval depends on the multiple alignment of all possible decoded bit data sequences from the genomic DNA. According to alignment-based data retrieval from the complete genome with multiple inserted sequences, data durability, error-correcting, and parity check operations are all supplemented without any other qualified DNA code or template DNA design.

Data Retrieval by Complete Genome Sequencing. Using this alignment-based approach, we demonstrated data storage in *B. subtilis* genomic DNA. For the secure retrieval of encoded data, our stratagem does not require additional material such as template DNA, only the sequencing of the complete genome. Data retrieval conducted by PCR-based amplification using the template DNAs designed for both the 5′- and 3′-regions of the single data region is vulnerable to the breakage of either of the DNA annealing sites, which are essential even to read the partial region of encoded data. Currently, a single DNA sequencer requires a 24-h day to sequence the equivalent of a bacterial genome, up to about two million bases (*12*). However, a growing demand for greater speeds and lower costs is pushing the development of new sequence technologies, and it is likely that new machines will be capable of reading one million bases per second in the relatively near future (*12*). This alignment-aided data storage and retrieval hence will be better suited to practical fast-lane genome sequencing techniques.

Preparation of Duplicated Data Cassettes. In order to recover data by sequence alignment, the length of data sequence cannot be shorter than a certain length, as it can otherwise occur by chance in the host genome. For instance, when a message data is by two keyboard inputs, our encryption key method creates an eight-nucleotide sequence within each cassette, which may have several possible alignment patterns at the genomic level. In such a case, to retrieve data securely we therefore need to add more redundant descriptions or contrive more complex codes that make longer data regions in each cassette. Although many bacteria species presumably contain a number of naturally repeating sequences, synthetic sequences of more than 20 nucleotides is a sufficiently specific feature ($p < 1 \times 10^{-5}$) in the 4.2 Mbp *B. subtilis* genomic DNA.

In order to prevent damage to genomic DNA during evolution, such as the deletion of junk DNA sequences especially caused by homologous recombination (*13*, *14*), our methodology introduced different nucleotide sequences encoding the same data by multiple data compression paths. Construction of a large piece of synthetic DNA is, however, still likely to be technically demanding and expensive. Although each nucleotide feature of the duplicated data sequence differed from each other and redundant descriptions were required to encode data, the direct data-to-nucleotide transitions maximized the information amount in each nucleotide cassette without any qualified DNA code, and the one-bit by one-bit frame shifting of the four-bit data windows generated the lossless data copies. Therefore, the cost of synthetic DNAs in this methodology is on a par with those of previous studies (*4*). The encryption keys of four-bit data to

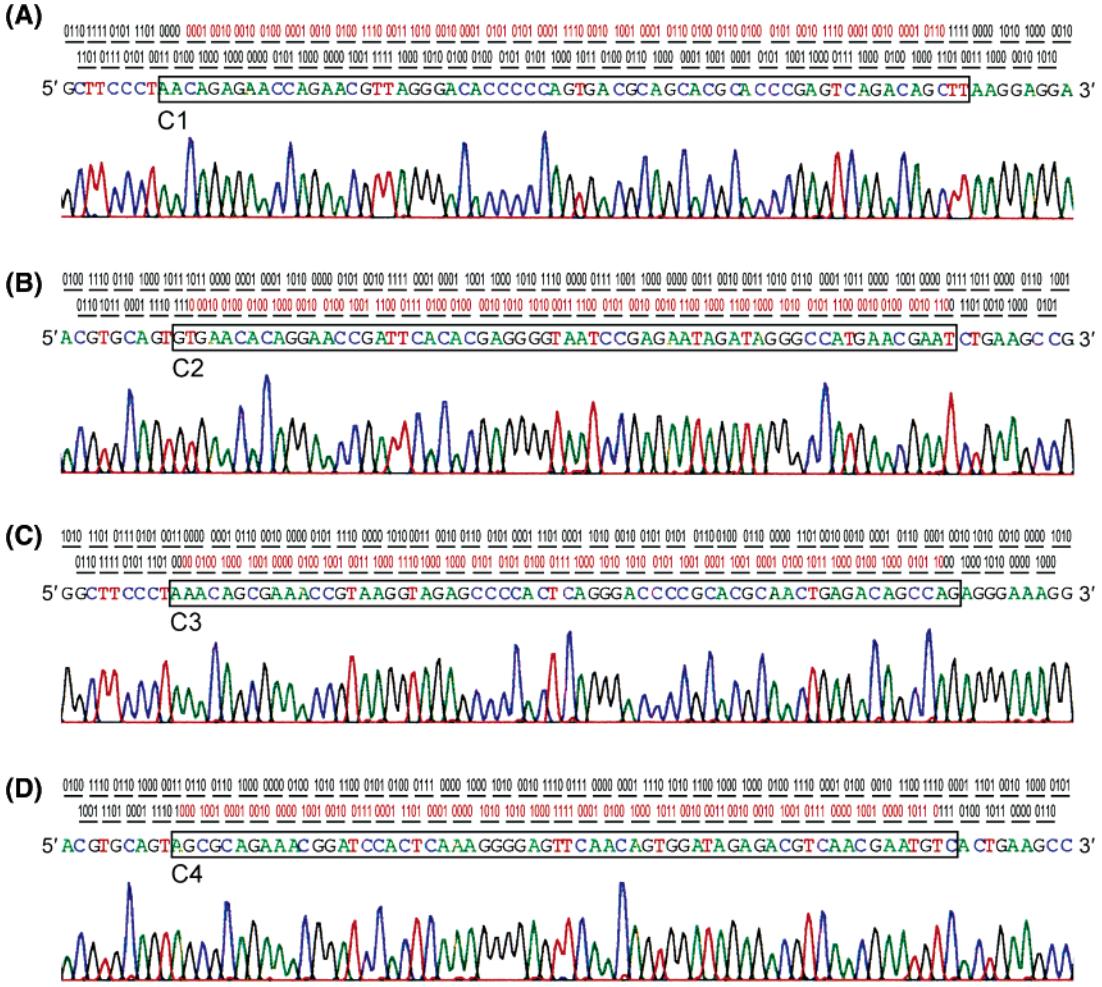

Figure 2. DNA sequencing of the four cassettes inserted into the *B. subtilis* genome. The genome sequences around the cassettes C1, C2, C3, and C4 (boxed) are shown in A, B, C, and D, respectively. According to the encryption keys, all possible dinucleotides were decompressed to the four-bit binary codes, which are displayed above each nucleotide sequence.

dinucleotides used in this study can replicate a maximum of four cassettes, but the number of cassettes is not limited by other encryption keys. For example, the encryption keys of eight-bit data to tetranucleotides have the potential to replicate data in eight cassettes.

Complementation of Durability by Alignment-Based Data Assembly. For data retrieval, genome sequences with multiple inserted nucleotide cassettes are decompressed to all possible patterns of bit data sequences. In this study, the encryption keys represent the one-to-one correspondence of dinucleotides and four-bit binary codes, and two of the bit data sequences are generated by a nucleotide sequence using two-bp sliding windows with one-bp displacement; the total of four possible bit data sequences are decompressed by the double-stranded genome sequence. Without any template DNA and parity check code within the bounds of data sequences, the encoded data and its sectional data breaks are recognizable by multiple alignments, searching conserved regions against all the decompressed bit data sequences. Point mismatch and gap space indicate mutations and partial deletions or insertions of genome sequence, respectively. The alignment result provides a complemented complete data region and is achieved by the merger of multiple decompressed patterns from the nucleotide fragments of the inserted cassettes. It complements frameshifts and adjusts encryption key reading frame misalignments. As such, it is similar to TBLASTX, an alignment program utilizing genetic codon rules that outputs possible amino acid sequences that are computationally translated from subjected nucleic acid sequences (*15*).

Although the multiplication of cassettes in this technique does lead to redundant volumes, data are retrievable independently of any other error-correcting algorithms, resulting in the advantage of data stability. Previous studies have revealed that intricately intertwined parity effects within single data-coding regions cost a certain volume of data sequence (*4*). The data recovery rate is fragile and proportional to data breakage, particularly that which occurs through long-range DNA deletion. For the approximation of data stability, we computationally simulated data retrievals from cassettes that were randomly mutated and deleted (Figure 3). In this simulation result, three or more cassettes were indicated to ensure a comparatively high recovery rate. Over 99% of the encoded data was recovered from four cassettes when their mutation/deletion rates were less than 15%. Moreover, the positions of the data breakages were easily identified by the alignment results when these data could not be rescued. The encoding of multiple data into separate sections, much like the redundant arrays of inexpensive disk (RAID) (*16*) technology frequently used in magnetic disk data storage, undoubtedly contributes to the stability and durability of the data. In addition, the absence of boundary characteristics from the template DNAs seen in previous studies enables the boundaries of the encoded sequences within genomic DNA to also be identified by alignment.

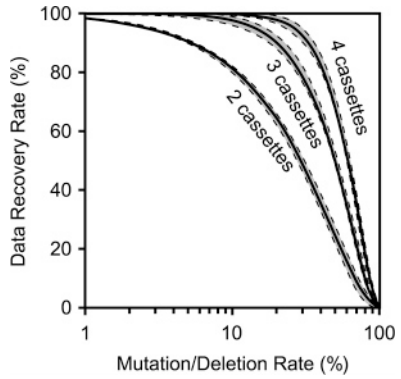

Figure 3. Expected data recovery rate. The solid lines represent the average data recovery rate per head of mutation/deletion. Dotted lines and gray-filled areas indicate two-sided 95% confidence intervals.

Data Storage into the *B. subtilis* Genomic DNA. Data storage into *B. subtilis* cells was performed by a two-step culture method. *B. subtilis* BEST2136 was transformed using pSPIBER01, generating GB01 (Table 2). Sequentially, GB01 was transformed using pSPIBER02, generating GB02 (Table 2). This strain harbors C1 and C2 at the *metB* locus and C3 and C4 at the *proB* locus. Insertion of the four cassettes was confirmed by DNA sequencing (Figure 2). GB02 is available upon request.

Practical Use of Long-Term and Large-Volume Data Storage. We have successfully demonstrated the data storage of the message "E=mc^2 1905!" into the *B. subtilis* strain BEST2136. One of the advantages of inserting data into organisms is the realization of data inheritance. Although these media vessels carry a significant risk of transmuting data in their evolutionary processes, our simple alignment-based approach effectuates a higher durability of data inheritance. While the size of the nucleotide sequence that can artificially be inserted into the organism's chromosomal DNA is limited (*3*), we suggest that *B. subtilis* is an acceptable species for large volume data storage. Previous megacloning techniques have cloned the entire 3.5 Mb genome of *Synechocystis* PCC6803 into the 4.2 Mb genome of *B. subtilis* 168, resulting in a 7.7 Mb composite genome (*17*). Hence, adopting and developing further codes and experimental methods or inserting plural fragments into the partial volumes of multiple-species metagenomes will enable the storage of huge volumes of data in heritable media. We suggest that this simple, flexible, and robust method offers a practical solution to data storage and retrieval challenges in combination with other, previously published techniques.

Acknowledgment

This work was inspired by discussions with Dr. Masanori Arita. The authors are grateful to the members of MGSP at the Institute for Advanced Biosciences, Keio University, for their critical discussions, Dr. Kazuharu Arakawa for helpful suggestions on the preparation of the manuscript, and Professor Mitsuhiro Itaya for providing *B. subtilis* strain BEST2136.

References and Notes

(1) Editorial, A Y3K bug. *Nat. Biotechnol.* **2000**, *18 (1)*, 1.

(2) Bancroft, C.; Bowler, T.; Bloom, B.; Clelland, C. T. Long-term storage of information in DNA. *Science* **2001**, *293*, 1763−1765.

(3) Cox, J. P. Long-term data storage in DNA. *Trends Biotechnol.* **2001**, *19*, 247−250.

(4) Smith, G. C.; Fiddes, C. C.; Hawkins, J. P.; Cox, J. P. Some possible codes for encrypting data in DNA. *Biotechnol. Lett.* **2003**, *25*, 1125−1130.

(5) Clelland, C. T.; Risca, V.; Bancroft, C. Hiding messages in DNA microdots. *Nature* **1999**, *399*, 533−534.

(6) Arita, M.; Ohashi, Y. Secret signatures inside genomic DNA. *Biotechnol. Prog.* **2004**, *20*, 1605−1607.

(7) Wong, P. C.; Wong, K.; Foote, H. Organic data memory using the DNA approach. *Commun. ACM* **2003**, *46*, 95−98.

(8) Arita, M. Comma-free design for DNA words. *Commun. ACM* **2004**, *47*, 99−100.

(9) Keyboard scan codes: Set 2. http://www.marjorie.de/ps2/scancode-set2.htm (accessed 7/23/06).

(10) Ohashi, Y.; Ohshima, H.; Tsuge, K.; Itaya, M. Far different levels of gene expression provided by an oriented cloning system in *Bacillus subtilis* and *Escherichia coli*. *FEMS Microbiol. Lett.* **2003**, *221*, 125−130.

(11) Dubnau, E.; Davidoff-Abelson, R. Fate of transforming DNA following uptake by competent *Bacillus subtilis*. I. Formation and properties of the donor-recipient complex. *J. Mol. Biol.* **1971**, *56*, 209−221.

(12) Bonetta, L. Genome sequencing in the fast lane. *Nat. Methods* **2006**, *3*, 141−147.

(13) Kowalczykowski, S. C.; Dixon, D. A.; Eggleston, A. K.; Lauder, S. D.; Rehrauer, W. M. Biochemistry of homologous recombination in *Escherichia coli*. *Microbiol. Rev.* **1994**, *58*, 401−465.

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

(15) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(16) Patterson, D.; Gibson, G.; Katz, R. A case for redundant arrays of inexpensive disks (RAID). *Proc. 1988 ACM SIGMOD Conf.* **1988**, *1*, 109−116.

(17) Itaya, M.; Tsuge, K.; Koizumi, M.; Fujita, K., Combining two genomes in one cell: Stable cloning of the *Synechocystis* PCC6803 genome in the *Bacillus subtilis* 168 genome. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102,* 15971−15976.

Received August 28, 2006. Accepted December 18, 2006.

BP060261Y