ORGANIC DATA MEMORY
Using the DNA Approach

For very long-term storage and retrieval, encode information as artificial DNA strands and insert into living hosts. As vectors, bacteria, even some bugs and weeds, might be good for hundreds of millions of years.

A DATA preservation problem looms over today's information superhighway. Ancient humans preserved their knowledge by engraving bones and stone. About two millennia ago people invented paper to publish their thoughts. Today, we use magnetic media and silicon chips to store our data. But bones and stone erode, paper disintegrates, and electronic memory degrades. All these storage media require constant attention to maintain their information content, and all are easily destroyed by people and natural disasters, whether intentionally or by accident. In light of the vast amount of information being generated every day, it's time to find a new medium.

Searching for an inexpensive, long-lasting medium for information storage, scientists at the Pacific Northwest National Laboratory (PNNL) are investigating deoxyribonucleic acid—commonly known as DNA—to develop a data memory technology with a life expectancy much greater than any existing counterpart. Our initial DNA memory prototype consists of four main steps: encoding meaningful information as artificial DNA sequences; transforming the sequences to living organisms; allowing the organisms to grow and multiply; and eventually extracting the information back from the organisms. Here we describe the objective of this investigation, which began in 1998, and experiments we've conducted to determine the feasibility of our approach, as well as several potential applications.

Nature magazine reported a study [1] resembling the first part of our effort—encoding meaningful information as DNA sequences. It described an experiment in which a group of scientists at Mount Sinai School of Medicine in New York created an encoded DNA strand and hid it behind a period (a dot) in a printed document. The document was then sealed and mailed to its owners through the U.S. Postal Service. Eventually, the embedded message was successfully recovered in a laboratory environment.

The article reported that the embedded information survived its rough handling in the mail, proving that a DNA strand can be as dependable as a piece of paper in terms of information storage. It is, however, still far from being able to outlast existing data-memory devices. In fact, a naked DNA molecule is easily destroyed in any open environment inhabited by people or potential enemies of nature. The so-called "double-strand break" of DNA, which is usually fatal, can be caused by common unfavorable environmental conditions, including excessive temperature and desiccation/rehydration. Even nucleases (a kind of DNA-degrading enzyme) in the environment can corrupt DNA molecules over time. Therefore, a key to our success is finding a super-dependable storage medium to ensure adequate protection for the encoded DNA strands. Our solution is to provide a living host for the DNA that tolerates the addition of artificial gene sequences and survives extreme environmental conditions. Perhaps more important, the host with the embedded information must be able to grow and multiply.

Challenges

Recent advances in genetic engineering have allowed the introduction of foreign DNA molecules into the living cells of bacteria, humans, and other organisms. Typically, a short, one-of-a-kind, well-researched DNA strand is applied to a living host for some particular biological study, with little or no intention of retrieving the embedded DNA afterward. This process is somewhat contrary to our basic DNA memory requirements that new and artificial DNA be generated frequently and that we be able to retrieve the embed-

BY PAK CHUNG WONG, KWONG-KWOK WONG, AND HARLAN FOOTE

ded information afterward.

These requirements pose serious challenges to our DNA memory design due to the size of a whole genome, which ranges from a few million DNA units in a bacterium to more than three billion in a mouse or human. It is practically impossible to retrieve an embedded message from a whole genome in a wet laboratory without knowing the content or whereabouts of the encoded DNA. The unpredictable nature of genomic mutation represents yet another obstacle, further reducing the odds of locating the message within a whole genome.

AAA - 0	AAC - I	AAG - 2	AAT - 3	ACA - 4	ACC - 5	ACG - 6	ACT - 7
AGA - 8	AGC - 9	AGG - A	AGT - B	ATA - C	ATC - D	ATG - E	ATT - F
CAA - G	CAC - H	CAG - I	CAT - J	CCA - K	CCC - L	CCG - M	CCT - N
CGA - O	CGC - P	CGG - Q	CGT - R	CTA - S	CTC - T	CTG - U	CTT -V
GAA - W	GAC - X	GAG - Y	GAT - Z	GCA - SP	GCC - :	GCG - ,	GCT - -
GGA - .	GGC - !	GGG - (	GGT -)	GTA - `	GTC -‘	GTG -“	GTT - "
TAA - ?	TAC - ;	TAG - /	TAT - [	TCA -]	TCC -	TCG -	TCT -
TGA -	TGC -	TGG -	TGT -	TTA -	TTC -	TTG -	TTT -

Table 1. DNA encoding.

Experimental Design

The customized computational and wet-laboratory approach we developed leaves a trail of the embedded message for later retrieval while allowing us to preserve the integrity of the message. Our experiments were carried out in four primary stages:

DNA host identification. In the process of identifying candidates to carry the embedded DNA molecules, we considered microorganisms (such as bacteria) and other agents (such as plants, including Arabidopsis) as message hosts. We eventually selected two well-understood bacteria—Escherichai coli (E. coli) and Deinococcus radiodurans (Deinococcus)—because microorganisms generally grow quickly and embedded information can be inherited quickly and continuously. We also considered the physical endurance of the DNA host candidates. Deinococcus, we learned, survive extreme conditions, including ultraviolet, desiccation, partial vacuum, and ionizing radiation up to 1.6 million rad, or radiation absorbed dose (about 0.1% of this dose is fatal to humans); some strains of Deinococcus also tolerate high temperatures.

Information encoding. The four basic building units in DNA are bases called deoxyribonucleosides. In the biology literature, they are usually labeled A (Adenine), C (Cytosine), G (Guanine), and T (Thymine). Each base always bonds with another base (such as A with T and C with G). A single chain of these bases is called an oligonucleotide, or oligo. It pairs with another complementary oligo (such as GATCG with CTAGC) to form a double-stranded DNA. Each of these AT or CG pairs in a double-stranded DNA is called a base pair. Because a DNA sequence is digital, we can use them to construct any English text, just as we use binary numbers 0 and 1 to encode ASCII characters. Table 1 outlines the encoding table of our experiments using a set of triplets—a DNA sequence with any three of the four bases—the exact encoding scheme for our initial experiment. Note several triplets listed at the end of the table are open (intentionally) for later expansion.

Unique DNA searching. The whole genomes of E. coli and Deinococcus have been completely sequenced and are available from The Institute for Genomic Research (www.tigr. org). Our task is to identify a set of fixed-size sequences (20-base-pair long in our experiments) that do not exist in the candidate bacteria yet satisfy all the genomic constraints and restrictions. This process is critical to our experiment, as we do not want to cause unnecessary mutation or damage to the bacteria. The resultant sequences also serve as sentinels to tag the beginning and end of the embedded messages—similar to file headers and footers in magnetic tape—for later identification and retrieval.

AAGGTAGGTAGGTTAGTTAG	AGAGTAGTGAGGATAGTTAG
AGGTTTGGTGGTATAGTTAG	ATAAGTAGTGGGGTAGTTAG
ATAGGAGTGTGTGTAGTTAG	ATAGGGGTATGGATAGTTAG
ATATTAGAGGGGGTAGTTAG	ATGGGTGGATTGATAGTTAG
GGAGTAGTGTGTATAGTTAG	GGGAATAGAGTGTTAGTTAG
GGGAGTATGTAGTTAGTTAG	GGGATGATTGGTTTAGTTAG
GGTTAGATGAGTGTAGTTAG	GTATGGGAATGGTTAGTTAG
TAAGGGATGTGTGTAGTTAG	TAGAGAGAGTGTGTAGTTAG
TAGAGGAGGGATATAGTTAG	TAGAGTGGTGTGTTAGTTAG
TAGATGGGAGGTATAGTTAG	TAGATTGGATGGGTAGTTAG
TAGGAGAGATGTGTAGTTAG	TAGGGTTGGTAGTTAGTTAG
TATAGGGAGGGTATAGTTAG	TATAGGGTAGGGTTAGTTAG
TGTGGGATAGTGATAGTTAG	

Table 2. 25 20-base-pair sequences for our experiment.

Of the 10 billion potential candidates in the bacterium Deinococcus, we found through intensive computation only 25 qualified sequences that would be acceptable for our experiments. These sequences (see Table 2) serve as blueprints for chemically synthesizing oligos for subsequent steps in our experiments. The multiple triplets (such as TAA, TGA, and TAG) seen in many of the sequences are called stop codons and tell the bacterium repeatedly it has reached the end of the native DNA sequence and should stop translating its contents. Without the protection of

stop codons, the bacterium could misinterpret the encoded information and produce artificial proteins that could destroy the integrity of the embedded message or even kill the bacterium.

Wet laboratory procedures. We conducted four main procedures:

Create complementary oligos. We started by creating two complementary oligos, each with 46 bases and consisting of two different segments of 20-base-long oligos connected by a six-base-long restriction enzyme site. The two 20-base-long oligos were based on two different sequences listed in Table 2. Enzymes that recognize a specific sequence of double-stranded DNA and that cut the DNA at that location are known as restriction enzymes. We created a restriction enzyme site for later insertion of encoded DNA fragments. These two 46-base-long complementary oligos form a double-stranded, 46-base-pair DNA fragment. The DNA fragment was then cloned into a recombinant plasmid—a union of foreign DNA fragments into a circular DNA molecule (see Figure 1). Because the two 20-base-long oligos do not exist in the genome of the host, they serve as identification markers for later message retrieval. The stop codons in these two oligos also help protect the message, as well as the host, from potential damage.

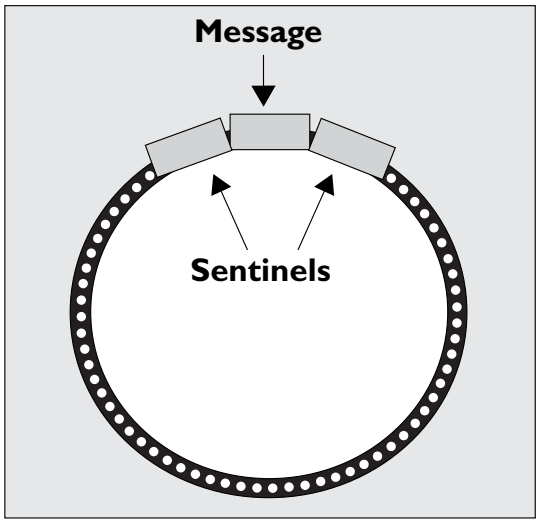

Figure 1. A recombinant plasmid with two DNA fragments as sentinels protecting the encoded message in between.

Insert DNA. The embedded DNA was then inserted into cloning vectors—a circular DNA molecule that can self-replicate within a bacterial host. The resultant vectors were then transferred into E. coli by electroporations (high-voltage shocks), allowing the vector with the encoded DNA fragment to multiply for later applications.

Incorporate into genome. The vector and the encoded DNA were then incorporated into the genome of Deinococcus for permanent information storage and retrieval. Deinococcus granted perfect protection for the embedded message, as it tolerates extreme desiccation, high doses of radiation, high levels of organic solvent, and vacuum-pressure environments, as shown in our experiments.

Extract the message. Finally, whenever embedded information was needed, we extracted the message part of the DNA strand from the bacterium through a laboratory procedure called polymerase chain reaction. Using prior knowledge of the sequences at both borders of the segments, it proceeds through a series of heating and cooling cycles to amplify the DNA segment. The whole process took about two hours. Figure 2 shows a machine readout of our DNA analysis and its English interpretation at the top.

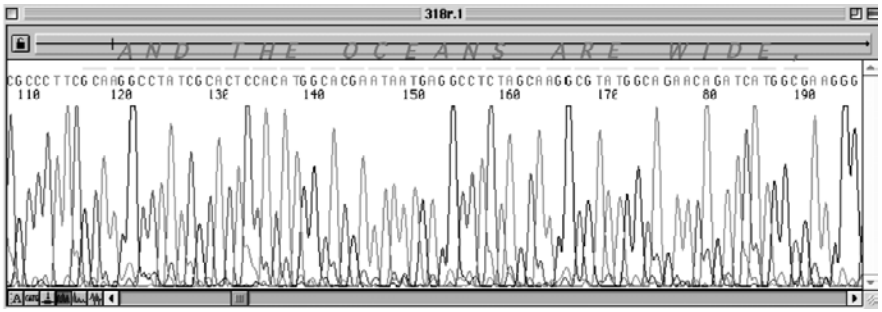

Figure 2. DNA analysis result and its interpretation; the English text is part of the children's song "It's a Small World" [2].

Enormous Potential Capacity

We successfully stored and retrieved seven chemically synthesized DNA fragments with 57–99 base pairs of non-native information in seven different individual bacteria. Even without further technology improvement, the capacity of bacterial-based DNA memory can be expanded dramatically by storing different pieces of information in a population of bacteria; for example, the seven bacteria in our experiment carry different parts of the children's song "It's a Small World" [2] in their genomes. Considering that a milliliter of liquid can contain up to 10^9 bacteria, the potential capacity of bacterial-based DNA memory is enormous, assuming we have a well-designed data index scheme.

A potential challenge is the mutation of the organisms affecting the integrity of the embedded messages. Although a bacterium can be selected with a low mutation rate, random changes still occur. Nature has had to deal with this problem since the beginning of bio-

logical evolution and developed mechanisms for detecting and correcting errors. With the extremely efficient DNA repair mechanisms associated with Deinococcus, we did not detect any mutations in our experiment in which we retrieved the DNA after the bacteria that carried the message was allowed to propagate for about a hundred generations. However, the mutation rate may depend on a specific sequence and the bacterium's genetic background.

DNA Memory Applications

Most of the potential applications of this organic data memory technology relate to the core missions of the U.S. Department of Energy (DOE). Other security-related applications include information hiding and data steganography for commercial products, as well as those related to national security.

As one of nine DOE national laboratories, a major PNNL concern is protecting information in case of nuclear catastrophe. Suppose the U.S. experienced a devastating nuclear disaster and the national information infrastructure was paralyzed or deactivated by radiation and fire. Suppose we had planted critical relief information in certain bacteria (such as Deinococcus) that could live and multiply independently without human intervention. Suppose these data hosts could survive high doses of radiation and other extreme conditions. All critical information would therefore be available upon the arrival of a disaster relief team.

The research into and development of sterile seeds—yield one crop, then terminate—has prompted recent controversy, especially in the farming community throughout the U.S. The competition between the proprietary rights of seed companies to protect their investments and the overwhelming need of poor farmers in third-world countries who cannot afford new seed every year will probably continue until a practical solution emerges. Suppose the seed companies were able to put unique DNA watermarks based on our technology in all their seeds. They could effectively track their sales and protect their proprietary products against illegal planting by greedy farmers without affecting the needy farmers.

Remediating environmental pollution in the U.S. has been a PNNL core mission since the 1980s. PNNL scientists periodically drill sampling wells and collect soil samples to monitor the migration of pollutants that might contaminate the U.S.'s natural resources, including water. Suppose we were able to put enough information in a bacteria population in the water and update it continuously and progressively according to the bacteria's spatial and temporal distribution. The bacteria would eventually provide both a chronological overview of the migration and a complete local database useful to scientists. The same technological approach could also be used to study endangered species; for example, a DNA watermark in the subject's genome could replace other artificial identification means (such as microchip implants).

For the computer science and information technology communities, suppose people could safely and permanently store their personal information (such as family history and medical data) in the cells in their own bodies. Suppose we could replace computer disks with our bodies as a primary memory storage medium.

Such options no longer represent speculative science fiction. All are potentially accomplished through organic data memory based on DNA. The DNA memory described here is neither impossible nor impractical—only challenging.

Conclusion

With a careful coding scheme and arrangement, important information can be encoded as an artificial DNA strand and safely and permanently stored in a living host. In the short run, this technology can be used to identify origins and protect R&D investment in, say, agricultural products and endangered species. It can also be used in environmental research to track generations of organisms and observe the ecological effect of pollutants. The microorganisms that survive heavy radiation exposure, high temperatures, and other extreme conditions are among the perfect protectors for the otherwise fragile DNA strands that preserve encoded information. Finally, living organisms, including weeds and cockroaches, that have lived on Earth for hundreds of millions of years represent excellent candidates for protecting critical information for future generations. C

References

1. Clelland Taylor, C., Risca, V., and Bancroft, C. Hiding messages in DNA microdots. *Nature 399* (June 10, 1999), 533–534.
2. Sherman, R.M. and Sherman, R.B. *It's a Small World.* Walt Disney Enterprises, Inc., 1963.

Pak Chung Wong (pak.wong@pnl.gov) is a chief scientist in the Energy Science and Technology Division at the Pacific Northwest National Laboratory, Richland, WA.

Kwong-Kwok Wong (kkwong@txccc.org) is an assistant professor at the Baylor School of Medicine and the Director of Microarray Laboratory at Texas Children's Cancer Center, Houston, TX. This research was conducted while he was a senior research scientist at the Pacific Northwest National Laboratory, Richland, WA.

Harlan Foote (harlan.foote@pnl.gov) is a senior research scientist in the Energy Science and Technology Division at the Pacific Northwest National Laboratory, Richland, WA.

The Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract DE-AC06-76RL0.