Biocompatible Writing of Data into DNA

Gary M. Skinner[1], Koen Visscher[1,2,3], and Masud Mansuripur[3]

[1] Department of Physics, The University of Arizona, Tucson, Arizona 85721
[2] Departments of Physics and Molecular and Cellular Biology, The University of Arizona, Tucson, Arizona 85721
[3] College of Optical Sciences, The University of Arizona, Tucson, Arizona 85721

[Published in the *Journal of Bionanoscience*, **Vol.1**, No.1, pp1-5 (2007). doi: 10.1166/jbns.2007.005]

Abstract. A simple DNA-based data storage scheme is demonstrated in which information is written using "address-encoded" oligonucleotides to encode two bits. In contrast to other methods that allow arbitrary code to be stored, the resulting DNA is suitable for downstream enzymatic and biological processing. This capability is crucial for DNA computers, and may allow for a diverse array of computational operations to be carried out using this DNA. Although here we use gel-based methods for information readout, we also propose more advanced methods involving protein/DNA complexes and atomic force microscopy/nano-pore schemes for data readout.

Introduction. DNA is attractive for storing digital information, partly due to its potentially ultra-high density; in theory, one gram of DNA is capable of storing about the same amount of data as 10^{12} CD-ROMs[1]. DNA also offers the possibility of creating extremely durable information archives, by introducing the DNA into a reproducing organism, such as bacteria tolerant to radioactivity[2]. As the organism replicates its genome, the information is carried into the next generation. In such a way, information can be secured for thousands, perhaps millions of years. The final key attraction of DNA is the novel forms of computational problems that can be addressed using it. For example, in his seminal paper, Adleman demonstrated that DNA could be used to solve an instance of the Hamiltonian path problem[3]. Effectively, DNA computing allows massively parallel searching of information space to find solutions to these kinds of problems.

Previous groups have encoded meaningful information directly into the sequence of base pairs[2,4]. However, chemical DNA synthesis is slow and expensive, and a new molecule must be created each time new data is to be written. To overcome this limitation, methods have been developed that take advantage of the formation of complementary base pairs in DNA to generate molecules representing information. Notably, a 3-bit system has been developed by Shin and Pierce that allows for the encoding of up to eight distinct states in a DNA molecule, and is even rewritable[5]. This method has some important limitations however, that may hinder its development into a fully fledged DNA memory device. It was not possible to unambiguously readout the information using simple gel-based methods, in this case only the total number of memory bits that were "1" or "0" could be determined[5]. To resolve the ambiguities, a fluorescence based system was included, using the quenching of different fluorophores at each memory location to determine the precise data content. This method is not practical for large capacity devices; a new fluorophore would be required for each new memory location, quickly exhausting all distinct fluorophore excitation/emission combinations available. A further limitation of this approach is that the resulting DNA construct is not easily manipulated by enzymatic or biological processes. Such a capability is critical, as the promise of DNA computers relies upon such manipulation of the encoded data to perform operations[3]. Specifically, this DNA construct is not linear, but resembles "frayed wires"[5], and if cloned into a living organism will not be faithfully replicated and also cannot be amplified using the polymerase chain reaction, as has been important in previous DNA computation demonstrations[3]. Although we do not claim that our method could be used directly in DNA computing, we hope that such ideas may be useful to those engaged in these endeavors.

In order to address these limitations we have developed an alternative DNA based memory that requires only simple methods to unambiguously readout the encoded information. Since the resulting molecules are linear double-stranded DNA, in principle this memory is compatible with a broad range of enzymatic and biological methods. Our method takes advantage of DNA recognition enzymes in order to distinguish between the "1" and "0" memory states. For this purpose we have chosen to use the restriction endonuclease EcoR1[6], an enzyme that binds to the palindromic sequence G^AATTC, cutting both DNA strands at the site marked "^". We chose this enzyme as it is well known that it possesses high stringency for its cognate recognition site[6]. By engineering the memory states "0" and "1" so that in one case the enzyme can recognize the DNA, and in the other it cannot, we can unambiguously determine the information content by a simple EcoR1 digestion procedure.

Our device operates on the principle of using linear DNA to specify a one dimensional array of writeable address locations (See Figure 1 and 3). Each of these locations is unique and the address space is large, for example by using 24 bases for the address, there are potentially 4^{24} unique address sequences. Each memory location consists of two parts, the addressing sequence, and the writeable memory sequence. The writeable sequence is composed of the recognition site for the restriction endonuclease, EcoR1, GAATTC. The addressing sequence is divided in half and flanks the EcoR1 site, this is done to allow sufficient space between each DNA site for the enzyme to bind and cut. To write to each memory location, an oligonucleotide that contains the complement to the addressing sequence is used. For each memory location in the device, two of these oligonucleotides are required, one to code for a "0" and one to code for a "1". These oligonucleotides differ only within the EcoR1 site itself, the region complementary to the memory sequence remains unchanged. The oligonucleotide that represents "1" contains a 4 nucleotide base mismatch, specifically GAATTC is changed to G***TTAA***C, such that the sequence will not base pair to the central four bases of the EcoR1 site in the device. The oligo for "0" is completely matched along the length of the memory element, and therefore the completely base-paired EcoR1 site is formed. The consequence of this is that if the mismatched oligonucleotide representing "1" is used to address the memory location, a double-stranded EcoR1 site cannot form and the enzyme will not bind to or cleave the DNA. The oligonucleotide representing "0" forms the complete EcoR1 site and therefore the protein can recognize its cognate site, and will cleave the DNA under normal conditions. In this demonstration we used DNA cleavage by the EcoR1 enzyme as the means to readout information from our device.

Methods and Results. In order to create a 2-bit capacity DNA based device with four distinct states (00, 01, 10, 11), five chemically synthesized ssDNA molecules were ordered from Integrated DNA Technologies; see Fig.1 for sequences. The first of these, "Blank_Medium" consists of 77 bases and contains two recognition sites for the Restriction Endonuclease EcoR1. These restriction sites are each flanked by regions of DNA serving as addressing sequences. For each location on "Blank_Medium", two shorter ssDNA addressing molecules were synthesized, for encoding "0" or "1", each containing the complement to the flanking addressing sequence. The "0" molecules also contain the complement to the EcoR1 sites, while the "1" molecules have four mismatched bases within the EcoR1 site, (GAATTC → G***TTAA***C).

To store information using these molecules, one need only mix together the "Blank_Medium" with either of the addressing molecules for each memory location, depending if a "0" or a "1" is desired. In practice, this was done by setting up a 50 µl annealing reaction in 1x NEB buffer #2 (10 mM Tris-HCl pH 7.9 @ 25°C, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM Dithiothreitol), with each ssDNA molecule present at 3 µM. The reaction was incubated at 95°C

to melt all base-pair interactions, and then slowly cooled to room temperature over several hours. Following this step, the information should then be stored within the molecules.

Subsequent readout of the stored digital information was achieved by performing a restriction digest with the enzyme EcoR1 (New England Biolabs). If a "0" was written at any given memory location, a completely base-paired cognate EcoR1 site is formed that will be recognized and cleaved by the enzyme. However, if a "1" was written, the four mismatched bases prevent the EcoR1 from recognizing and cleaving the DNA. By subsequent gel electrophoresis of the digestion products it is now possible to unambiguously determine the information encoded. In practice, 10 µl of the encoded product DNA was added to 40 µl NEB buffer #2 plus 40 Units of EcoR1. These reactions were incubated for 48 hours at 37 °C to ensure complete digestion of all EcoR1 sites. The reaction products were then analyzed by polyacrylamide gel electrophoresis (PAGE) in Tris-borate EDTA buffer at 10V/cm. A 10 bp DNA step ladder was run alongside the samples in order to determine fragment lengths on the gel (Fig.2).

Following EcoR1 digestion, the resulting products were analyzed using a native polyacrylamide gel; see Fig.2. As can be seen, each state of the DNA memory device is easily distinguished from the other three and the expected DNA fragments are present as described above. Each of the four states "00", "01", "10" and "11" has been faithfully recorded into the DNA and can be unambiguously read using EcoR1, as expected.

Discussion. The presented DNA memory device is easily scaled to much larger capacities; one need only increase the overall length of the "Blank_Medium" molecule to include additional memory locations and synthesize the corresponding addressing molecules for the new locations. It should be noted that these molecules must be made, in the first instance, by chemical synthesis. However, following this initial step, any arbitrary data can be stored without the need for synthesis of further molecules. In this way, our DNA memory device is a truly generic storage medium.

In this prototype device, the EcoRI sites were engineered so as to yield differently sized fragments upon digestion, e.g. to distinguish "01" from "10". As the capacity of such a memory device increases, this method for removing ambiguity will become increasingly impractical. In a larger device, the sites should be equally spaced along the molecule and the data stored as a palindrome within the DNA, i.e. all memory locations duplicated and reflected around the center of the molecule (Figure 3a). With such a construct all states will yield a unique pattern of digestion products. This has the added advantage of introducing redundancy into the memory, while decreasing the information density by only a factor of two.

The resulting DNA memory is ready for use in further enzymatic manipulations; the fact that EcoR1 is able to digest the product efficiently already shows that it is a suitable substrate for at least this class of DNA processing enzymes. By sealing the single-stranded "nicks" in the DNA between adjacent memory locations, using ligase, it should be possible to amplify such DNA by PCR, and potentially to introduce it into living systems; such steps have been shown to be crucial for successful DNA information technology applications[2-4]. It would be true that only 50% of the PCR amplified DNA would be of the sequence desired, but by molecular cloning of these sequences into single-copy plasmids (e.g., Novagen's pETcoco-1), the clones could be screened and only the desired sequence selected from resulting transformants.

We do not expect that restriction digestion and gel analysis will be the method of choice for readout of the data in a practical application, especially when the capacity is increased. The gel method used here has two distinct limitations, firstly it is slow, and very long gels would have to

be run to resolve the digestion fragments in larger devices. Furthermore, as the data is read in this fashion it is destroyed. However, we propose an alternative readout strategy that would overcome these limitations. A mutant version of EcoR1 exists that has the ability to bind tightly to the cognate DNA site, but is unable to cut the molecule[7]. It simply remains bound, forming a protein "stud" on the DNA; such mutant EcoR1/DNA complexes studded along a length of λ-DNA have been imaged using atomic force microscopy (AFM) [7]. It should be possible to use a similar method to decorate our memory device with EcoR1, each binding to the cognate sites located at the "0" positions, while leaving gaps where the EcoR1 is unable to bind the mismatched sites that represent "1". The readout could be performed using AFM and subsequent image analysis, as was done with EcoR1 on λ-DNA[7]. However, a more exciting prospect is to feed the protein/DNA complex through a solid-state nanopore[8] forming a "Nanoscale digital tape drive" (Figure 3b). The pore diameter would be engineered to yield a large current blockade differential when EcoR1/DNA complexes pass through as opposed to when naked DNA strands pass through. The palindromic nature of the encoded data means that there is no need to control the orientation of the DNA molecule as it enters the nanopore. Thus, the data could be read as a sequence of blockade events, and the intact protein/DNA complex could then be returned to a storage area to be accessed later.

By having a robust, simple means to encode a sequence of digital information into DNA in a form that is suitable for a wide range of further manipulations, the goal of practical, large scale DNA computing may be a step closer to reality.

Acknowledgement. This work is supported by the Office of Naval Research MURI grant No. N00014-03-1-0793, G.M.S. is a fellow of the Jane Coffin Childs Memorial Fund for Medical Research (www.jccfund.org) and the University of Arizona BIO5 Institute.

References

1. Kashiwamura, S.; Yamamoto, M.; Kameda, A.; Shiba, T.; Ohuchi, A., Potential for enlarging DNA memory: the validity of experimental operations of scaled-up nested primer molecular memory. *Biosystems* **2005,** 80, (1), 99-112.
2. Wong, P. C.; Wong, K. K.; Foote, H., Organic data memory using the DNA approach. *Communications of the Acm* **2003,** 46, (1), 95-98.
3. Adleman, L. M., Molecular computation of solutions to combinatorial problems. *Science* **1994,** 266, (5187), 1021-1024.
4. Clelland, C. T.; Risca, V.; Bancroft, C., Hiding messages in DNA microdots. *Nature* **1999,** 399, (6736), 533-4.
5. Shin, J. S.; Pierce, N. A., Rewritable memory by controllable nanopatterning of DNA. *Nano Letters* **2004,** 4, (5), 905-909.
6. McClarin, J. A.; Frederick, C. A.; Wang, B. C.; Greene, P.; Boyer, H. W.; Grable, J.; Rosenberg, J. M., Structure of the DNA-Eco RI endonuclease recognition complex at 3 A resolution. *Science* **1986,** 234, (4783), 1526-41.
7. Allison, D. P.; Kerper, P. S.; Doktycz, M. J.; Spain, J. A.; Modrich, P.; Larimer, F. W.; Thundat, T.; Warmack, R. J., Direct atomic force microscope imaging of EcoRI endonuclease site specifically bound to plasmid DNA molecules. *PNAS* **1996,** 93, (17), 8826-8829.
8. Storm, A. J.; Storm, C.; Chen, J.; Zandbergen, H.; Joanny, J. F.; Dekker, C., Fast DNA translocation through a solid-state nanopore. *Nano Lett* **2005,** 5, (7), 1193-7.

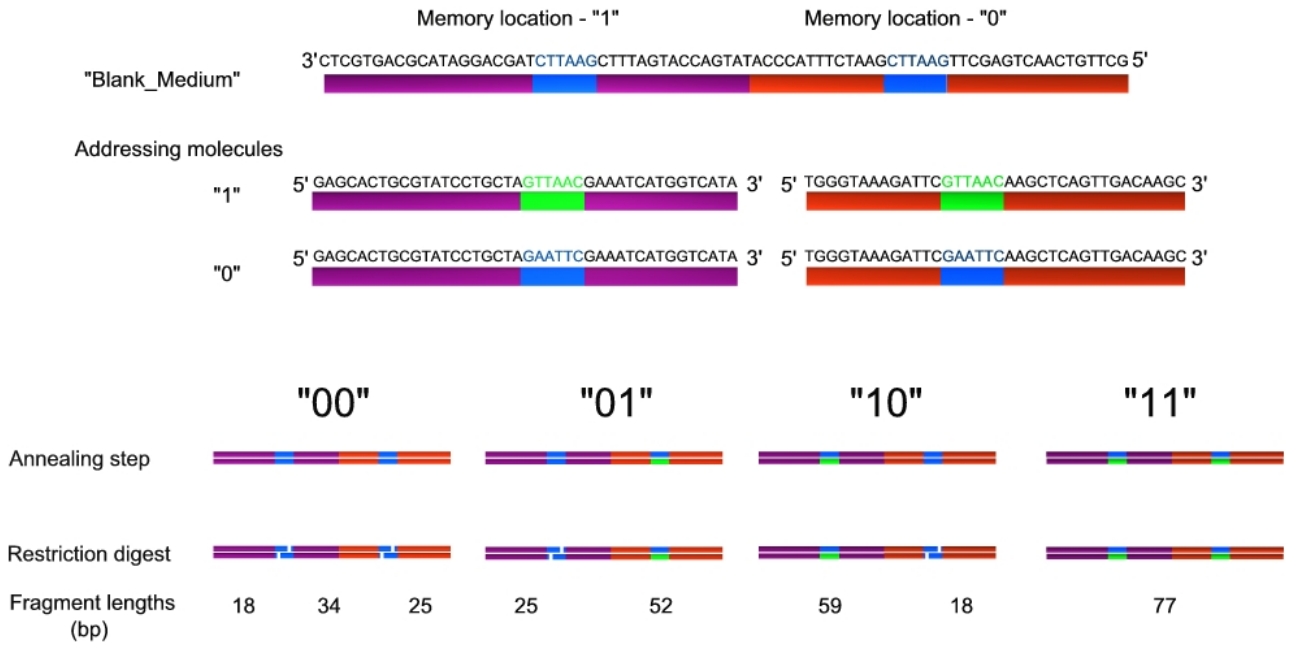

Figure 1. Memory sequences used are shown at the top, with the EcoRI sites (Blue), and the sequences comprising memory location 1 (Magenta) and memory location 2 (Red). In the address-encoded molecules, the binary digit "1" is represented by a partially mismatched EcoRI site (Green), while the binary digit "0" has the full complement to the EcoRI site (Blue). Shown are 2-bit sequences (00, 01, 10, 11) written onto blank media. Prior to annealing, the blank medium and specific addressing molecules are added to the reaction in order to achieve the desired data sequence after annealing. Shown are the expected DNA fragments resulting from EcoR1 digestion of the encoded DNA products.

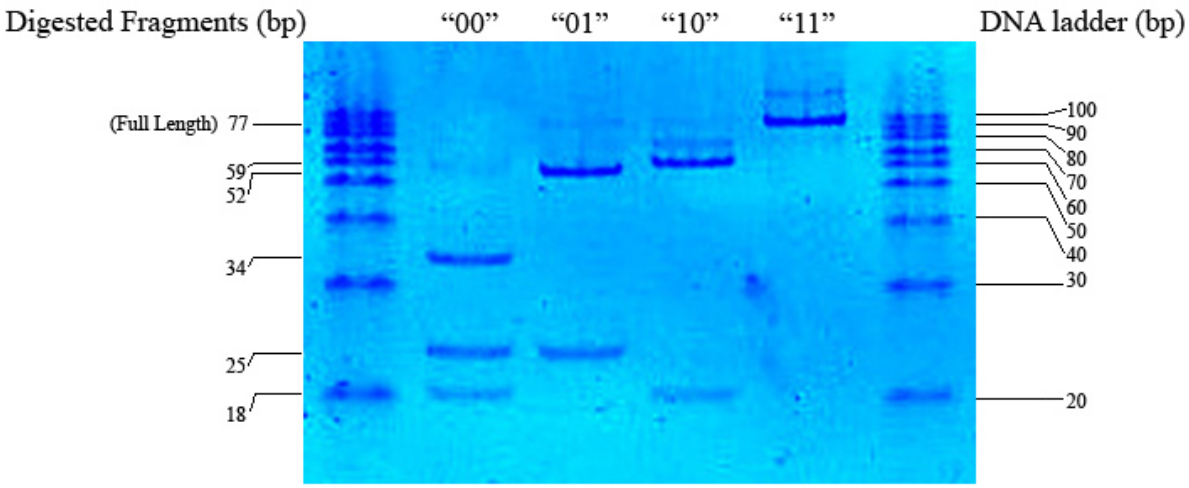

Figure 2. Restriction digest with EcoRI enzyme yields a distinct pattern of DNA fragments**,** which depends upon the specific EcoRI sites that have been protected. In this way the four states of the 2-bit device, 00, 01, 10 and 11 can unambiguously be determined.

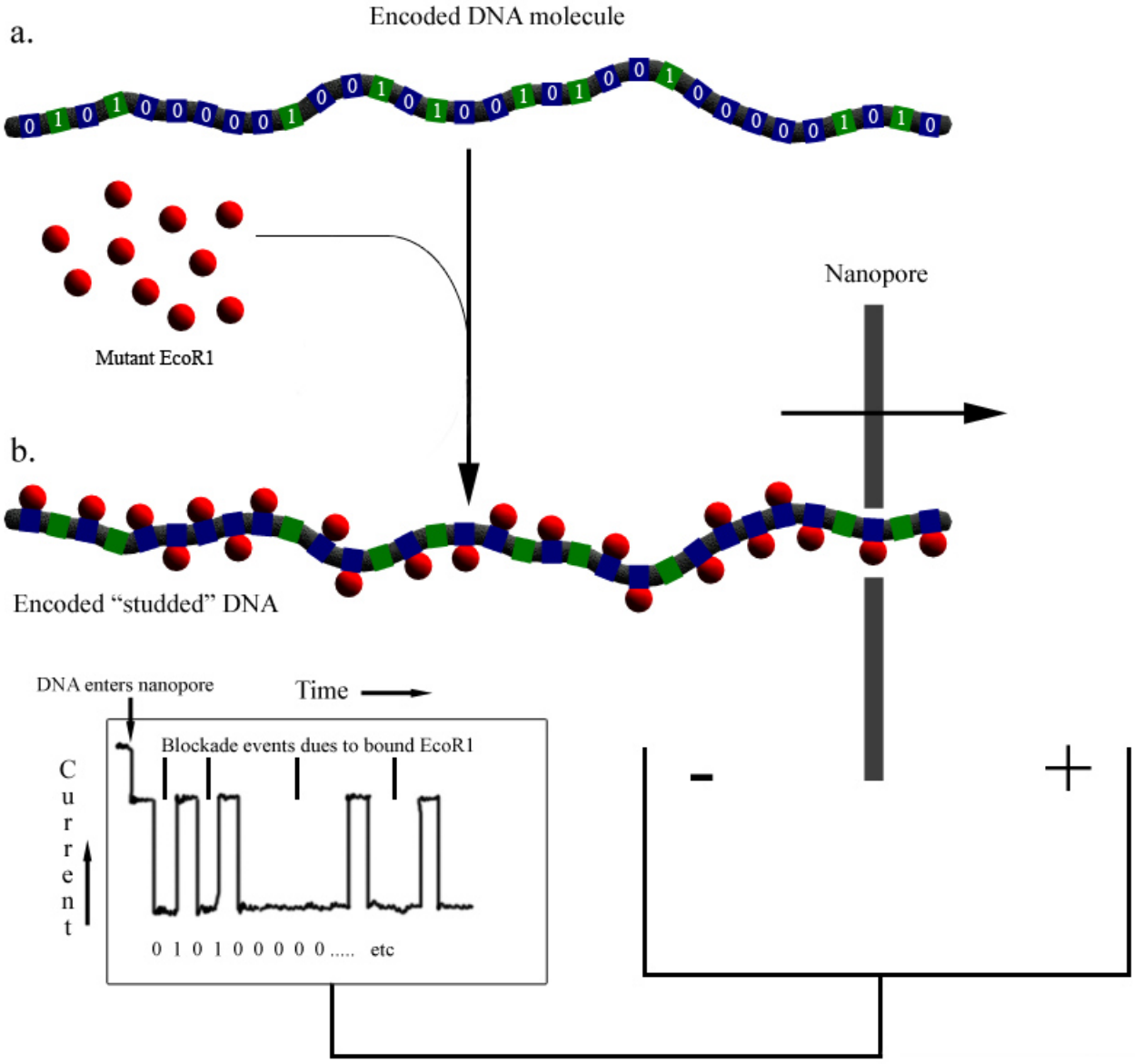

Figure 3. Advanced readout system. (a) The encoded DNA molecule is incubated with the mutant form of EcoR1 that is able to bind tightly to its correct DNA site but is not able to cleave the DNA[7]. (b) The resulting "studded" DNA is then fed through a solid-state nanopore, driven by an applied voltage. The current through the nanopore is measured as a function of time and discrete additional blockade events would be observed (simulated here) each time an EcoR1/DNA complex passes through the nanopore. This current versus time data can then be used to determine the digital sequence of the encoded DNA.