Length-Based Encoding of Binary Data in DNA

Nathaniel G. Portney,[†] Yonghui Wu,[‡] Lauren K. Quezada,[#] Stefano Lonardi,[†] and Mihrimah Ozkan*[,§,||,⊥]

Departments of Bioengineering, Computer Science and Engineering, Biochemistry, and Electrical Engineering, and Center for Nanoscale Science and Engineering, University of California, Riverside, California 92521, and Department of Biology, Loyola Marymount University, 1 LMU Drive, Los Angeles, California 90045

Received October 18, 2007. In Final Form: December 18, 2007

We developed a system to encode digital information in DNA polymers based on the partial restriction digest (PRD). Our encoding method relies on the length of the fragments obtained by the PRD rather than the actual content of the nucleotide sequence, thus eliminating the need for expensive sequencing machinery. In this letter, we report on the encoding of 12 bits of data in a DNA fragment of 110 nucleotides and the process of recovering the data.

Dexyribonucleic acid (DNA) has a promising future in the area of digital information storage because of its high capacity, stability, and error resilience (e.g., see Bancroft[1] and Cox[2] for examples of DNA digital storage prototypes). DNA-based storage has already spawned a variety of commercial applications. For example, several companies provide DNA-based technologies to prevent and identify counterfeits (see, e.g., Applied DNA Sciences at http://www.adnas.com/, PSA/DNA at http://www.psadna.com/, and DNA Technologies at http://www.dnatechnologies.com/). The object that needs to be authenticated is labeled with a tiny quantity of DNA that cannot be easily detected. The detection can be achieved by PCR amplification, provided that one knows a primer (which plays the role of a cryptographic key).

The most straightforward way to store information in DNA is to use one nucleotide base to encode two bits of information.[1,2,4] Here we report on an alternative method to store information in DNA that takes advantage of the partial restriction digest (PRD) for the decoding. In our scheme, restriction enzyme sites are dispersed in a DNA sequence in a particular pattern according to the actual information to be stored. The information is retrieved by PRD, followed by gel electrophoresis.

Our scheme offers several advantages over previous methods.[1,2,4] First, the decoding in our method is simpler and cheaper because it does not require sequencing, which in turns needs expensive equipment. Second, it allows one to choose the content of the nucleotide bases that are not occupied by the restriction enzyme sites, thus allowing the construction of more stable DNA sequences (e.g., one can attempt to avoid the formation of secondary structures or achieve a specific GC content).

Figure 1 illustrates the process by which DNA decoding was achieved. Initially, a plasmid containing the information-sensitive “decoding fragment” (DF) is excised by flanking Stu I restriction sites using complete digestion, followed by 5′ hot labeling by ^{32}P. The 110 bp DF contains rich Alu I restriction sites spaced every 4 bp or 8 bp to encode a 1 or a 0, respectively. Alu I partial restriction digestion (PRD) that is used to obtain all possible 5′-labeled fragments are then separated by electrophoresis, following band decoding to yield the final message “MEMO” according to the algorithm described below.

Transfected bacterial plasmid expression vector (custom pJ5, DNA2.0) with 110 bp decoding fragment (DF) sequence 5′-AGG|CCTTTGTTGTGGTTGGTGTTGTGGTGGTAG|-CTATCGAG|CTAG|CTTAGCAG|CTAG|CTAG|CTAG|-CTAG|CTAG|CTCGATAG|CTAG|CTCGTAAG|CTAG|CTAG|-CTAG|CTAAAAGG|CCT-3′ was designed to utilize Stu I (AGG|CCT) restriction sites flanking our desired DF sequence (5′-CCT...AGG-3′). A series of 4 bp (CTAG) and 8 bp (CTTAGCAG) repeating fragments separated by Alu I (AG|CT) sites were carefully positioned to produce a 1 or 0 binary code according to our decoding algorithm. A 4 bp difference between decodable fragments (4 or 8) was selected for conservative gel-resolution limitations. A clonally derived DF method was used over chemical synthesis and PCR because of poor fidelity from the statistical termination of oligos.[5] Cell culturing of bacterially transfected pJ5 containing DF was cultured on agar (Invitrogen, ImMedia Kan Agar cat. no. Q611-20), and a picked colony was inoculated in LB broth containing Kanamycin (Invitrogen, ImMedia Kan Liquid cat. no. Q610-20) for 18 h at 37 °C with shaking to amplify the DF insert. Plasmid isolation preparation was performed (Qiagen QIAfilter Plasmid Mega, cat. no. 12281) to isolate the pJ5 plasmid with $A_{280\ \mathrm{nm}}/A_{260\ \mathrm{nm}} = 1.9$. Stu I complete digestion (NEB, cat. no. R0187S) of pJ5 was performed (27.63 U/μg template h, 37 °C) for 1 h to isolate DF (Figure 2), followed by 20 min of thermal deactivation at 65 °C. Column purification (Qiagen MinElute Reaction Cleanup, cat. no. 28204) of excised DF was used to eliminate enzyme and impurities. Manual denaturing TBE-Urea PAGE (12.5%, 5 h, 136 V) was stained with GelGreen (Biotium CA, cat. no. 41004) intercalating dye and imaged on a Typhoon 9410 imaging system ($\lambda_{ex} = 488$ nm, PMT 700, 200 μm resolution) with a 560LP emission filter. Figure 2 PAGE results show the 2277 bp pJ5 plasmid and 110 bp DF following Stu I complete digestion.

The purified DF product (79 μg/μL, $A_{260}/A_{280} = 2.0$) was dialyzed for 30 min with 1 L of DI dialysate on 26.34 μL of

* Corresponding author: E-mail: mihri@ee.ucr.edu. Phone: (951)827-2900. Fax: (951)827-2525.

[†] Department of Bioengineering, University of California.

[‡] Department of Computer Science and Engineering, University of California.

[§] Department of Biochemistry, University of California.

[||] Department of Electrical Engineering, University of California.

[⊥] Center for Nanoscale Science and Engineering, University of California.

[#] Loyola Marymount University.

[1] Bancroft, C.; Bowler, T.; Bloom, B.; Clelland, C. T. *Science* **2001**, 293, 1763−1765.

[2] Cox, J. P. L. *Trends Biotechnol.* **2001**, *19*, 247−250.

[3] Adleman, L. M. *Science* **1994**, *266*, 1021−1024.

[4] Chen, J.; Deaton, R.; Wang, Y. *DNA Comput.* **2004**, *2943*, 145−156.

[5] Hecker, K. H.; Rill, R. L. *BioTechniques* **1998**, *24*, 256−260.

[6] Nishigaki, K.; Kaneko, Y.; Wakuda, H.; Husimi, Y.; Tanaka, T. *Nucleic Acids Res.* **1985**, *13*, 5447−5760.

10.1021/la703235y CCC: $40.75 © xxxx American Chemical Society
Published on Web 01/30/2008 PAGE EST: 3.4

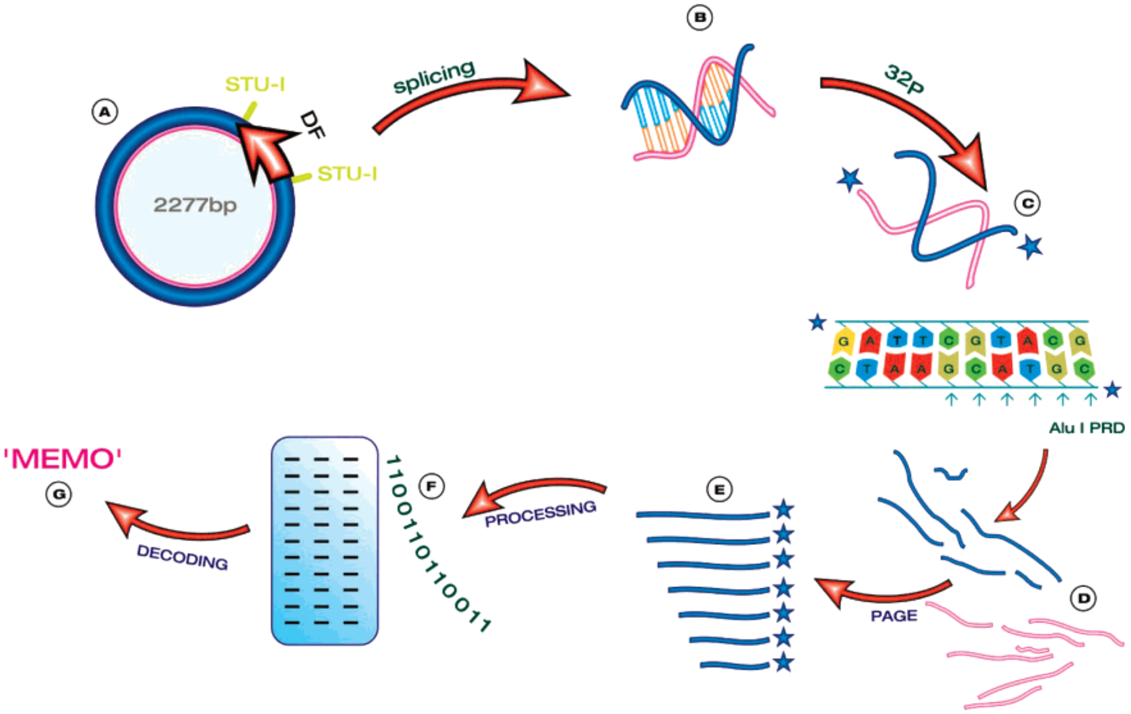

Figure 1. Illustration of the length-based decoding process. (A) The plasmid expression vector is used to amplify the decoding fragment (DF) at the cloning site (red arrow), which is spliced by endonuclease Stu I at each end and purified in B. ^{32}P labeling using AP and T4 PnK labels for the 5′ of the DF fragment to achieve C. Distinctly repeated Alu I recognition sites are partially digested to fragment the DF in D and are separated by denaturing PAGE in E. Processing of the gel F by or decoding algorithm allows the conversion of the band pattern to binary code, demonstrating the use of length-based DNA to decode the final message MEMO in G.

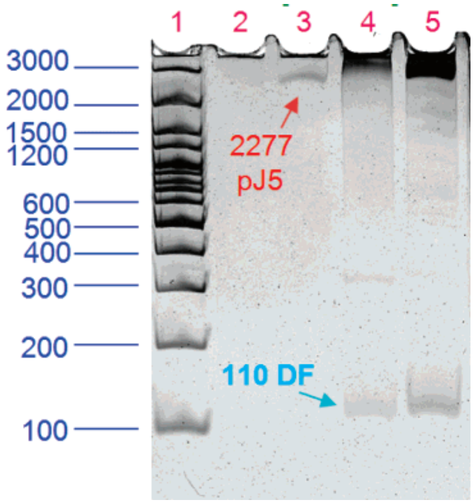

Figure 2. Isolation of decoding fragment (DF). Denaturing PAGE (12.5%, 136 V, 5 h) showing the isolation of pJ5 (L3) and the excision of 110 bp DF by Stu I endonuclease (L4, L5) alongside marker L1. Fluorescence imaging was provided by intercalating the GelGreen 30 min stain (488 nm excitation, 560LP emission filter, PMT 700) on a Typhoon 9410 imaging system (PMT 700, 100 μm resolution).

product. The dephosphorylation procedure to remove 5′ phosphate on DF was performed by alkaline phosphatase reaction (AP, NEB no. M0290S) with 5 U/μg at 37 °C for 15 min, following heat deactivation for 5 min at 65 °C. Prior to being hot labeled, the substrate was heated for 10 min at 70 °C before adding 10 U/μL □-^{32}P (80 μCi) with T4 polynucleotide kinase (NEB T4 PnK, cat. no. M0201S) and incubating for 30 min at 37 °C, followed by heat deactivation at 70 °C for 20 min. A serial dilution for the Alu I PRD reaction on radiolabeled DF was performed using 0, 0.1, 1, 10, 50, and 100 U of Alu I/μg h at 37 °C. Alu I RE was chosen for duplex-independent activity, a short, symmetric recognition sequence, thermal deactivation, nonvariable splicing, and reasonable C-G content. Incubation at 37 °C for 1 h was followed by thermal deactivation at 65 °C for 20 min. Similar 5′ hot probe labeling of BsuRI-digested pBR322 (Fermentas, cat. no. SM0271) was used as a marker. A 6% manual denaturing PAGE (3 h, 1200V) was run with hot-labeled Alu I PRD products exposed to an X-ray film for 48 h and imaged on a Typhoon 9410 imaging system (Figure 4).

The double-stranded DF design contains a 30 bp-length precursor region (5′-CCTTTGTTGTGGTTGGTGTTGTGGTG-GTAG-3′) on the leading strand, which is 5′ radiolabeled and not divisible by the 4 or 8 bp internal fragments. The Alu I PRD procedure produces all possible labeled fragments {30, 38, 42, 50, 54, 58, 62, 66, 70, 78, 82, 90, 94, 98, 102, 110}. Because T4 PnK also labels the 5′ on the reverse complement DF, an equivalent, dependent reverse complement band fragment distribution {8, 12, 16, 20, 28, 32, 40, 44, 48, 52, 56, 60, 68, 72, 80, 110} is produced. For example, a 42 bp fragment on a forward strand method is equivalent to 68 bp (110-42) on a reverse strand method, by the conserved 110 bp DF length. The decoding algorithm first determines whether each band is from the forward or reverse strand. If a band is from the forward strand, then it is converted to the corresponding reverse strand. Then, the bands are sorted, and differences between consecutive bands are computed. If the difference is 4 nucleotides, then bit 1 is produced; if the difference is 8, then bit 0 is produced.

Multimode data fitting was used to linearly correlate unknown PRD fragments with a calibration marker and a band pattern software tool developed by the authors. According to the gel in Figure 4, the 20 bands of BsuR1 digested the pBR322 standard (Fermentas, cat. no. SM0271) with a 15 bp custom marker, which was found to appear as three natural fitting regimes, as represented in the Figure 3 calibration plot. This three-mode linear semilog fitting produced an R^2 of up to 0.99. A total of 13 observed {8,

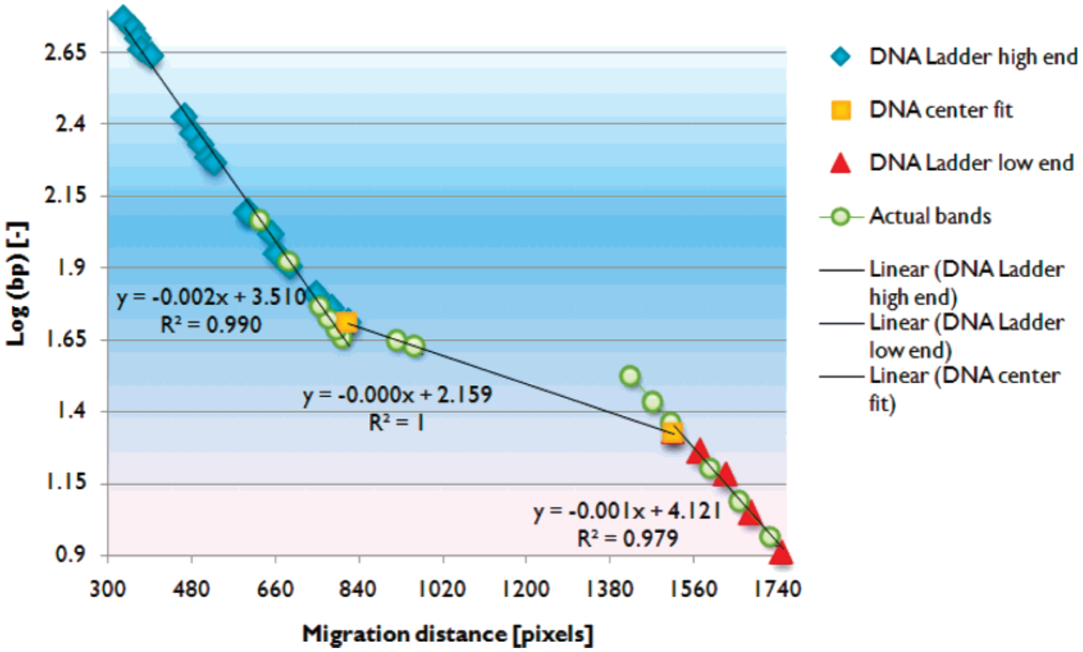

Figure 3. DNA fitting for decoding analysis. Semilog plot of $\log_{10}$(bp) vs the migration distance in pixels, which reveals distinct modes of fitting. Three-mode fitting of the pBR322 DNA ladder from P32 labeling PAGE shows calibration regions for the high (blue diamonds), middle (orange squares), and low (red triangles) ranges with a correlation value of $R^2 \approx 1$. Unknown decoding band positions (13 or 15) in green collapsed on the calibration plot. Decoding bands reveal two calculated bands decoded within the middle fitting region, five out of seven decoded within the high fitting region, and all six bands decoded in the low fitting region. The average error calculated unknown relative to the computed unknown is 1.06 bp.

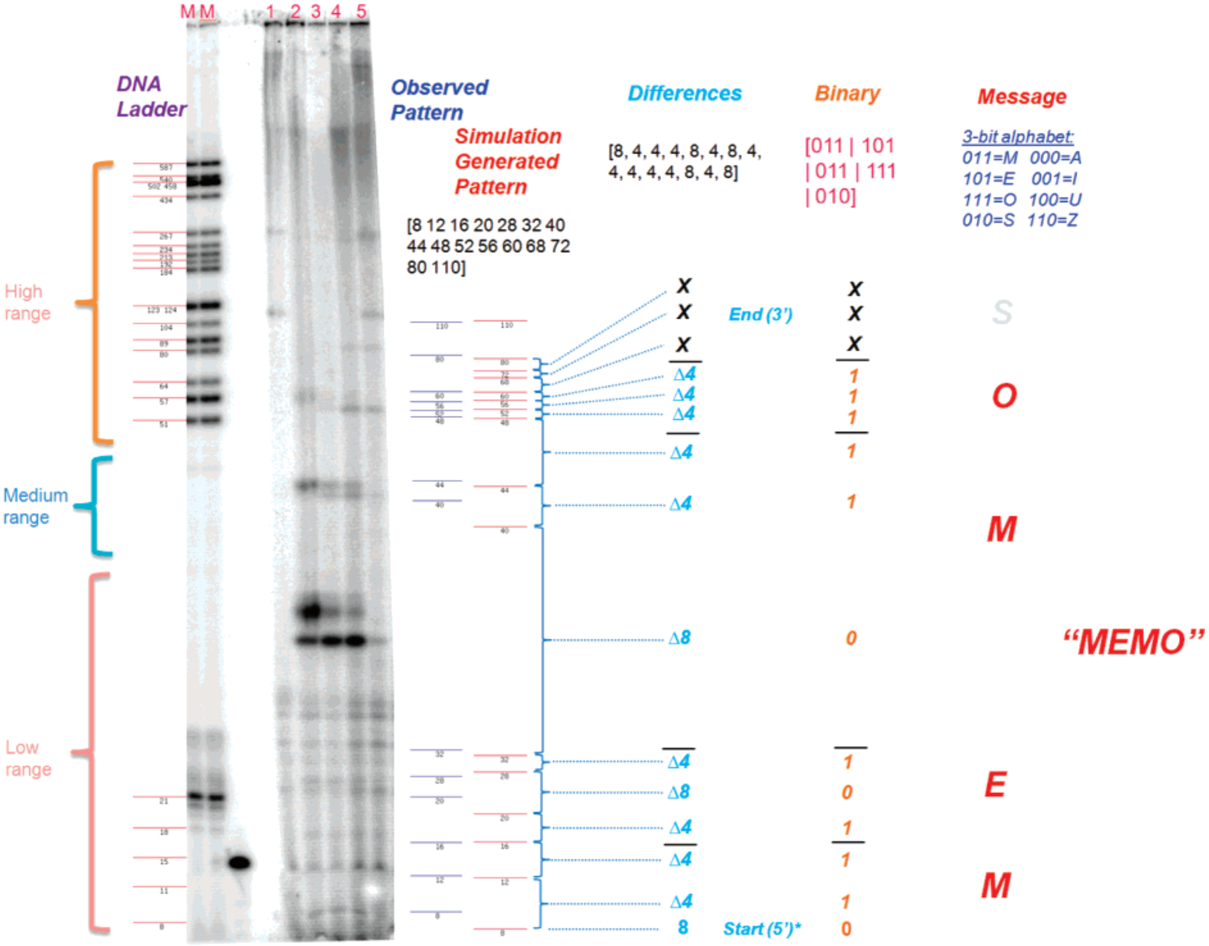

Figure 4. Decoding of message from electrophoresis. (A) P32-labeled denaturing PAGE of the P32-labeled pBR322 DNA marker (M lanes), with the hot-labeled Alu I partially digested decoding fragment (increasing enzyme from L1-L5). (B) Simulated band pattern from the marker lane is consistent with the observed band pattern. Two bands at 68 and 72 are missing from the observed band pattern. Differences of 8 or 4 bp from the band pattern are converted to binary 1 or 0 using a reverse strand designation. Using a 3-bit alphabet, each consecutive 3 binary digits yields to a letter to form the final message MEMO with the final S being absent.

12, 16, 20, 28, 32, 40, 44, 48, 52, 56, 60, 80} relevant decoding band positions out of the 15 expected {8, 12, 16, 20, 28, 32, 40, 44, 48, 52, 56, 60, 68, 72, 80} were correlated using this calibration to quantify the band lengths of unknowns within 2 bp accuracy. The software tool developed to produce an image of expected band patterns alongside the experimentally observed pattern of appropriate lengths uses a GD library (http://www.libgd.org/) as its underlying image-rendering tool. This software is available upon request from the authors. Using this tool, calculated band patterns from decoding relevant bands (dismissing internal bands) were located.

The decoding of unknown bands (Figure 4) was performed using the aforementioned multimode fitting analysis to locate expected band positions relative to a generated simulation pattern and the application of our decoding algorithm. Internal bands are dismissed because they are not terminally labeled. The 3-bit alphabet was used {A = 000, E = 101, I = 001, M = 011, O = 111, S = 010, U = 100, Z = 110}, applying a decoder function that first transforms a band from the forward strand to an equivalent band from the reverse band. Then the bands from the reverse strand are sorted according to their length. The differences between successive bands are calculated. If the difference is 8, it encodes a 0, and if the difference is 4, it encodes a 1. The selection in this manner maximizes the encoding capacity per nucleotide. For example, to achieve the first letter M in our scheme (Figure 4), the successive binary code 011 must be obtained, which requires consecutive band differences of 8, 4, and 4. Using calibration fitting to quantify final band positions, the smallest fragment starts at 8 bp (Δ8 bp), with the next lengths being 12 bp (Δ4 bp) and 16 bp (Δ4 bp). Therefore, Δ8 bp, Δ4 bp, and Δ4 bp achieve 011 and the letter M letter according to our 3-bit alphabet. Continuing to read off trios of binary code from the 5′ bottom position reveals a full binary code (011|101|011|111|010) from band differences of 844|484|844|444|848 to produce MEMOS. The final S was not realized because of the absence of bands 68 and 72 bp, thereby shortening the final message to MEMO, albeit preserving 87% decoding accuracy.

We reported on a method to store binary in DNA that exploits the partial restriction enzyme digest rather than sequencing. Our method requires a rather inexpensive procedure that may be used to decode sensitive data in the field using a dual digest following the electrophoresis procedure. Although the storage density is just 0.11 bits/nucleotide, the decoding process dismisses sequencing completely. By using a single gel with 24 lanes, one could resolve 288 bits of data in several hours with only femtomolar quantities of material.

Acknowledgment. We thank Dr. Yinsheng Wang for the use of laboratory space and Jaclyn Kaladjian,Yong Jiang, and Alfredo Martinez for technical assistance. N.G.P. and M.O. are grateful for financial support from the DOE-DARPA CNID center and the FCRP-DARPA FENA center. S.L. and Y.W. were supported in part by NSF Career grant IIS-0447773 and NSF grant DBI-0321756.

LA703235Y