Sciencexpress

Brevia

Next-Generation Digital Information Storage in DNA

George M. Church,[1,2] Yuan Gao,[3] Sriram Kosuri[1,2]*

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [2]Wyss Institute for Biologically Inspired Engineering, Boston, MA 02115, USA. [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA.

*To whom correspondence should be addressed. E-mail: sri.kosuri@wyss.harvard.edu

As digital information continues to accumulate, higher density and longer-term storage solutions are necessary (*1*). DNA has many potential advantages as a medium for immutable, high latency information storage needs (*2*). For example, DNA storage is very dense. At theoretical maximum, DNA can encode two bits per nucleotide (nt) or 455 exabytes per gram of ssDNA (*3*). Unlike most digital storage media, DNA storage is not restricted to a planar layer, and is often readable despite degradation in non-ideal conditions over millennia (*4*, *5*). Finally, DNA's essential biological role provides access to natural reading and writing enzymes and ensures that DNA will remain a readable standard for the foreseeable future.

Storing messages in DNA was first demonstrated in 1988 (*6*) and the largest project to date encoded 7920 bits (*7*). The small scale of previous work stems from the difficulty of writing and reading long perfect DNA sequences, and has limited broader applications (table S1). Here, we develop a strategy to encode arbitrary digital information using a novel encoding scheme that utilizes next-generation DNA synthesis and sequencing technologies (fig. S1). We converted an html-coded draft of a book that included 53,426 words, 11 JPG images and 1 JavaScript program into a 5.27 megabit bitstream (*3*). We then encoded these bits onto 54,898 159nt oligonucleotides (oligos) each encoding a 96-bit data block (96nt), a 19-bit address specifying the location of the data block in the bit stream (19nt), and flanking 22nt common sequences for amplification and sequencing. The oligo library was synthesized by ink-jet printed, high-fidelity DNA microchips (*8*). To read the encoded book, we amplified the library by limited-cycle PCR and then sequenced on a single lane of an Illumina HiSeq. We joined overlapping paired-end 100nt reads to reduce the effect of sequencing error (*9*). Then using only reads that gave the expected 115-nt length and perfect barcode sequences, we generated consensus at each base of each data block at an average of ~3000-fold coverage (fig S2). All data blocks were recovered with a total of 10 bit errors out of 5.27 million (table S2), which were predominantly located within homo-polymer runs at the end of the oligo where we only had single sequence coverage (*3*).

Our method has at least five advantages over past DNA storage approaches. We encode one bit per base (A or C for zero, G or T for one), instead of two. This allows us to encode messages many ways in order to avoid sequences that are difficult to read or write such as extreme GC content, repeats, or secondary structure. By splitting the bit stream into addressed data blocks, we eliminate the need for long DNA constructs that are difficult to assemble at this scale. To avoid cloning and sequence verifying constructs, we synthesize, store, and sequence many copies of each individual oligo. Since errors in synthesis and sequencing are rarely coincident, each molecular copy corrects errors in the other copies. We use a purely in vitro approach that avoids cloning and stability issues of in vivo approaches. Finally, we leverage next-generation technologies in both DNA synthesis and sequencing to allow for encoding and decoding of large amounts of information for ~100,000-fold less cost than first generation encodings.

The density (5.5 petabits/mm^3 at 100x synthetic coverage) and scale (5.27 megabits) of this work compare favorably to other experimental storage technologies while only using commercially available materials and instruments (Fig. 1 and table S3). DNA is particularly suitable for immutable, high-latency, sequential access applications such as archival storage. Density, stability, and energy efficiency are all potential advantages of DNA storage (*10*), while costs and times for writing and reading are currently impractical for all but century-scale archives (*3*). However, the cost of DNA synthesis and sequencing have been dropping at exponential rates of 5- and 12-fold per year, respectively – much faster than electronic media at 1.6-fold per year (*11*). Hand-held, single-molecule DNA sequencers are becoming available, and would vastly simplify reading DNA-encoded information (*12*). Our general approach of using addressed data blocks combined with library synthesis and consensus sequencing should be compatible with future DNA sequencing and synthesis technologies. Reciprocally, large-scale use of DNA such as for information storage could accelerate development of synthesis and sequencing technologies (*13*). Future work could use compression, redundant encodings, parity checks, and error correction to improve density, error rate, and safety. Other polymers or DNA modifications can also be considered to maximize reading, writing, and storage capabilities (*14*).

References and Notes

1. "Extracting Value from Chaos" (IDC, Framingham, MA, 2011), http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.
2. C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland, Long-term storage of information in DNA. *Science* **293**, 1763 (2001). doi:10.1126/science.293.5536.1763c Medline
3. Information on materials and methods is available on *Science* Online.
4. J. Bonnet *et al.*, Chain and conformation stability of solid-state DNA: Implications for room temperature storage. *Nucleic Acids Res.* **38**, 1531 (2010). doi:10.1093/nar/gkp1060 Medline
5. S. Pääbo *et al.*, Genetic analyses from ancient DNA. *Annu. Rev. Genet.* **38**, 645 (2004). doi:10.1146/annurev.genet.37.110801.143214 Medline
6. J. Davis, Microvenus. *Art J.* **55**, 70 (1996). doi:10.2307/777811
7. D. G. Gibson *et al.*, Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52 (2010). doi:10.1126/science.1190719 Medline
8. E. M. LeProust *et al.*, Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522 (2010). doi:10.1093/nar/gkq163 Medline
9. J. St. John, *SeqPrep* https://github.com/jstjohn/SeqPrep (2011)
10. L. M. Adleman, Molecular computation of solutions to combinatorial problems. *Science* **266**, 1021 (1994). doi:10.1126/science.7973651 Medline
11. P. A. Carr, G. M. Church, Genome engineering. *Nat. Biotechnol.* **27**, 1151 (2009). doi:10.1038/nbt.1590 Medline
12. E. Pennisi, Search for pore-fection. *Science* **336**, 534 (2012). doi:10.1126/science.336.6081.534 Medline
13. S. Kosuri, A. M. Sismour, When it rains, it pores. *ACS Synth. Biol.* **1**, 109 (2012). doi:10.1021/sb300015f
14. S. A. Benner, Z. Yang, F. Chen, Synthetic biology, tinkering biology, and artificial biology. What are we learning? *C. R. Chim.* **14**, 372 (2011). doi:10.1016/j.crci.2010.06.013
15. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012). doi:10.1038/nmeth.1923 Medline
16. H. Li *et al*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009). doi:10.1093/bioinformatics/btp352 Medline
17. C. T. Clelland, V. Risca, C. Bancroft, Hiding messages in DNA microdots.

Nature **399**, 533 (1999). doi:10.1038/21092 Medline

18. P. Wong, K. Wong, H. Foote, Organic data memory using the DNA approach. *Commun. ACM* **46**, 95 (2003). doi:10.1145/602421.602426
19. C. Gustafsson, For anyone who ever said there's no such thing as a poetic gene. *Nature* **458**, 703 (2009). doi:10.1038/458703a
20. N. Yachie, K. Sekiyama, J. Sugahara, Y. Ohashi, M. Tomita, Alignment-based approach for durable data storage into living organisms. *Biotechnol. Prog.* **23**, 501 (2007). doi:10.1021/bp060261y Medline
21. N. G. Portney, Y. Wu, L. K. Quezada, S. Lonardi, M. Ozkan, Length-based encoding of binary data in DNA. *Langmuir The Acs Journal Of Surfaces And Colloids* **24**, 1613 (2008). doi:10.1021/la703235y Medline
22. M. Ailenberg, O. Rotstein, An improved Huffman coding method for archiving text, images, and music characters in DNA. *Biotechniques* **47**, 747 (2009). doi:10.2144/000113218 Medline
23. Ecma International, *Data interchange on read-only 120mm optical data disks (CD-ROM)*, (ECMA Standard 130, Geneva, Switzerland 1996, http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-130.pdf).
24. Ecma International, *120 mm DVD - Read-Only Disk*, (ECMA Standard 267, Geneva, Switzerland 2001, http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-267.pdf).
25. Blu-Ray Disc Association, *White Paper – Blu-Ray Disc Format* (2nd Edition, Universal City, CA 2010, http://www.blu-raydisc.com/Assets/Downloadablefile/general_bluraydiscformat-15263.pdf).
26. Oracle, *StorageTek T10000 Family Tape* Cartridge (Oracle, Redwood Shores, CA 2010, http://www.oracle.com/us/products/servers-storage/storage/tape-storage/033617.pdf).
27. SanDisk, *SanDisk Develops Smallest 128Gb NAND Flash Memory Chip* (SanDisk, Milipitas, CA 2012, http://www.sandisk.com/about-sandisk/press-room/press-releases/2012/sandisk-develops-worlds-smallest-128gb-nand-flash-memory-chip).
28. Toshiba, *NAND Flash Memory in Multi Chip Package* (Toshiba, Tokyo, Japan, 2011, http://www.toshiba-components.com/memory/mcp.html).
29. Seagate, *Seagate Reaches 1 Terabit Per Square Inch Milestone In Hard Drive Storage With New Technology Demonstration* (Seagate, Cupertino, CA, 2012, http://www.seagate.com/about/newsroom/press-releases/terabit-milestone-storage-seagate-master-pr/).
30. S. Loth, S. Baumann, C. P. Lutz, D. M. Eigler, A. J. Heinrich, Bistability in atomic-scale antiferromagnets. *Science* **335**, 196 (2012). doi:10.1126/science.1214131
31. D. M. Eigler, E. K. Schweizer, Positioning single atoms with a scanning tunnelling microscope. *Nature* **344**, 524 (1990). doi:10.1038/344524a0
32. C. R. Moon, L. S. Mattos, B. K. Foster, G. Zeltzer, H. C. Manoharan, Quantum holographic encoding in a two-dimensional electron gas. *Nat. Nanotechnol.* **4**, 167 (2009). doi:10.1038/nnano.2008.415 Medline
33. T. Grotjohann *et al.*, Diffraction-unlimited all-optical imaging and writing with a photochromic GFP. *Nature* **478**, 204 (2011). doi:10.1038/nature10497 Medline
34. H. E. Kubitschek, Cell volume increase in *Escherichia coli* after shifts to richer media. *J. Bacteriol.* **172**, 94 (1990). Medline
35. "Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA" *Federal Registrar* **75**, 62820-62832 (2010) FR Doc No: 2010-25728.

Acknowledgments: This work was supported by U.S. Office of Naval Research (N000141010144), Agilent Technologies, and the Wyss Institute. Agilent Technologies is a commercial provider for DNA microchips. G.M.C. and S.K. designed and performed all experiments and analyses, and wrote the manuscript; Y.G. performed the sequencing. We thank J. Aach, C. Fracchia, S. Raman, H. H. Wang, A. W. Briggs, J. Lee, T. Wu, and D. B. Goodman for helpful suggestions on the manuscript.

Supplementary Materials
www.sciencemag.org/cgi/content/full/science.1226355/DC1
Materials and Methods
Supplementary Text
Figs. S1 and S2
Tables S1 to S3
References (*15*–*35*)

20 June 2012; accepted 07 August 2012
Published online 16 August 2012
10.1126/science.1226355

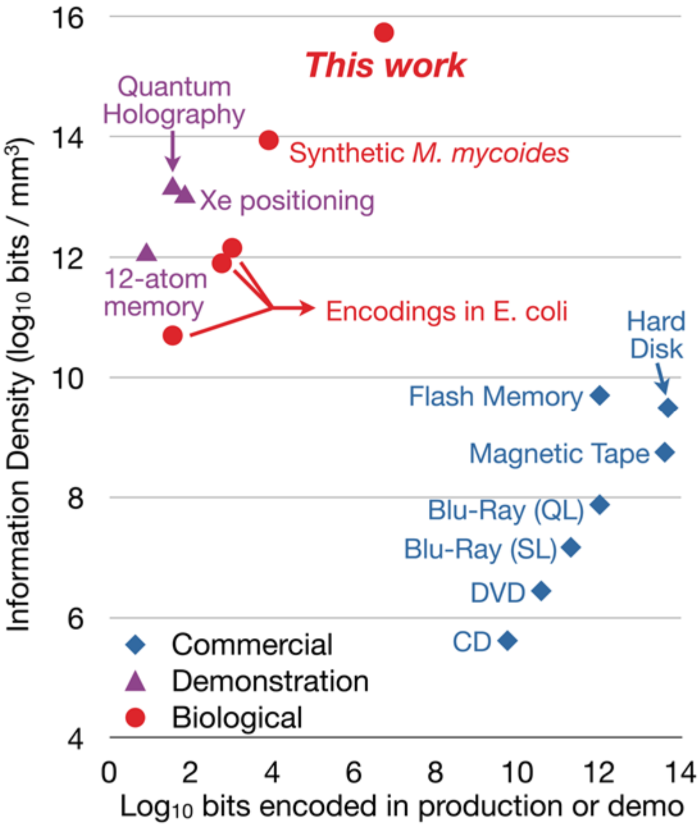

Fig. 1. Comparison to other measured by the $\log_{10}$ of bits encoded in the report or commercial technologies. We plotted information density ($\log_{10}$ of bits/mm^3) versus current scalability as unit (*3*).