Bioengineered 4:3, 123–125; May/June 2013; © 2013 Landes Bioscience

Synthetic DNA
The next generation of big data storage

Aisling O' Driscoll[1] and Roy D. Sleator[2,*]

[1]Department of Computing; Cork Institute of Technology; Cork, Ireland; [2]Department of Biological Sciences; Cork Institute of Technology; Cork, Ireland

With world wide data predicted to exceed 40 trillion gigabytes by 2020, big data storage is a very real and escalating problem. Herein, we discuss the utility of synthetic DNA as a robust and eco-friendly archival data storage solution of the future.

We are now in the Century of Biology[1] and in this new era, the petabyte (PB) is the currency. According to the International Data Corporation (IDC), it is estimated that worldwide data, approximated at 0.8 ZB (a trillion GB) in 2009, will increase to 40 ZB by 2020.[2] In light of this, solutions such as cloud computing have been proposed as the savior of storage, with the cloud storage market alone projected to pass $46 billion by 2018. However, to quote Einstein, "We can't solve problems by using the same kind of thinking we used when we created them." Therefore, the key to our data storage problems may lie not in thinking bigger but rather in thinking smaller.

> "Our remedies oft
> in ourselves do lie"
> Shakespeare, *All's Well That Ends Well* (I, i, 231–232)

According to papers published recently in Science[3] and Nature[4] by researchers at Harvard and EMBL-EBI, respectively, DNA, the original information storage molecule comprising the biological script of life, may hold the solution to our future data storage problems. DNA is a high-capacity storage medium, with a theoretical storage potential of 455 exabytes per gram ssDNA.[3] As a consequence, all of the world's projected 40 ZB of data could be stored in ~90 g of DNA. In addition to this, molecular biology now provides us with the tools to cut (restriction endonucleases[5,6]), paste (DNA ligase[7,8]) and copy (PCR[9]) DNA as we might the text of a word document. Furthermore, DNA is an extremely stable molecule, with a remarkably long life-span even in suboptimal environments, making it an ideal archival material.[10] Indeed, more than 80% of the woolly mammoth (*Mammuthus primigenius*) genome, comprising 3.3 billion nt, remains readable despite the fact that this species disappeared from the planet at the end of the Pleistocene (10,000 y ago).[11] Such nuclear genome sequencing of extinct species reveals population differences not evident from the fossil records and has even led to the discovery of genetic factors that may have affected species extinction.[12]

Some of the first attempts to use DNA as a workable canvas for archival purposes include Joe Davis' Microvenus; a 35-bit coded visual icon representing the external female genitalia and by coincidence, an ancient Germanic rune representing the female Earth.[13] More recently, construction of JCVI-syn1.0,[14] the first bacterial cell to contain a completely synthetic genome, employed "watermarks" to distinguish the synthetic genome from native DNA. These 7,920-bit watermarks contain strings of bases that, in code, spell out a web address, the names of the paper's authors and quotations ascribed to Joyce, Oppenheimer and Richard Feynman.[15] Although successful on a small scale, a significant limitation to the large scale practical application of DNA-based information storage is the difficulty of synthesizing long stretches of DNA de novo. Church and colleagues[3] at Harvard were the first to attempt to overcome these difficulties using next-generation DNA synthesis and sequencing technologies. Rather than a single long stretch of DNA (representing the complete data string), the team opted to work with shorter, overlapping fragments which together contain all the necessary information, yet individually are easier to manipulate in vitro. Furthermore, in order to move beyond the limited encoding of uppercase text which was the basis of previous approaches, the Harvard team chose to code an entire book (*Regenesis: How Synthetic Biology Will Reinvent Nature and Ourselves* ISBN-13:978-0465021758), including 53,426 words, 11 JPG images (at 10:1 data compression) and one JavaScript program. The team began by converting the text to html format using the Universal Character Set Transformation Format, 8-bit (UTF-8), backward compatible with ASCII and UNICODE for special fonts and character sets. The html-coded draft was then converted into a 5.27-megabit bitstream with the resulting bit sequence subsequently converted to DNA code using a 1-bit per base encoding (A,C = 0; T,G = 1), disallowing homopolymer runs greater than three while balancing GC content. The 5.27-megabit bitstream encoded 54,898 oligos, each 159 nt in length and consisting of a 96-bit data block (96 nt), a 19-bit address (19 nt) specifying the data block location and flanking 22 nt common sequences to facilitate amplification and sequencing. Following limited cycle PCR, to amplify the library, the sequence was read using an Illumina HiSeq next generation sequencer. With ~3,000-fold nucleotide coverage, all data blocks were recovered with a total of 10 bit errors out of

*Correspondence to: Roy D. Sleator; Email: roy.sleator@cit.ie
Submitted: 03/13/13; Accepted: 03/13/13
http://dx.doi.org/10.4161/bioe.24296

Figure 1. DNA-based data storage—the big data storage solution of the future?

5.27 million (most of the errors being predominantly located within homopolymer runs and at the sequence ends with only single sequence coverage).

In an effort to improve upon Church's work, Goldman et al.[4] recently described a modified strategy, which seeks to significantly reduce error and, as a result, facilitate up-scaling of DNA-based data storage. Achieving a storage density of ~2.2 PB/g DNA (equivalent to ~468,000 DVDs), the Goldman et al.[4] approach first converts the original file type to binary code (0, 1) which is then converted to a ternary code (0, 1, 2), which is in turn converted to the triplet DNA code. Replacing each trit with one of the three nucleotides different from the preceding one (i.e., A, T or C, if the preceding one is G) ensures that no homopolymers are generated—significantly reducing high throughput sequencing errors.[16] A further error limiting strategy involved the generation of overlapping segments (100 nt long data blocks with 75 nt overlap; alternate segments being converted to their reverse complement), creating 4-fold redundancy. Given that a majority of the errors associated with the Church method can be ascribed to either lack of coverage and/or homopolymers (runs of ≥ 2 identical nucleotides), the increased redundancy and lack of homopolymers of the Goldman et al.[4] strategy means that it is significantly less error prone than its predecessor. As proof of concept, the authors targeted four different file types (totaling 739 kilobytes of hard-disk storage):

- ASCII: the text file of a compression algorithm, Huffman code and all 154 of Shakespeare's sonnets
- PDF: the classic 1953 Watson and Crick[17] DNA structure paper
- JPEG: a color photograph of the authors' host institution, the European Bioinformatics Institute
- MP3: a 26 sec excerpt from Dr Martin Luther King, Jr's "I have a dream" speech

In line with the approach taken by Church and colleagues, all five files were represented by short stretches of DNA, specifically 153,335 strings, each comprising 117 nt (incorporating both data and address blocks to facilitate file determination and localization within the overall data stream). The oligos were synthesized using Agilent's oligo library synthesis process (creating ~1.2×10^7 copies of each DNA string), before being read using an Illumina HiSeq sequencer. Four of the five files were fully decoded without intervention (the fifth contained two 25 nt gaps which were easily closed following manual inspection), resulting in overall file reconstruction at 100% accuracy.

Based on a fixed string length (data and indexing) of 117 nt, Goldman et al.[4] suggest that DNA-based storage currently remains feasible even at several orders of magnitude greater than current global data volumes (measured in the ZB scale, 10^{21} bytes). This, combined with the likely expectation of significantly longer string synthesis as the technology progresses,[18] virtually future proofs DNA as a viable storage medium. Despite this, cost still remains an important limiting factor. Current costs, estimated to be in the order of €12,400/MB of storage, are impractical for all but century-scale archives, with limited access requirements. However, if a similar exponential correlation between storage space and cost is experienced, as was the case over the past 40 y [a 1 GB (1,000 MB) hard drive costing ca. $1,000,000 in 1980 is now available for less than 10 cents] and given the decline in DNA synthesis and sequencing costs (dropping at a rate of 5- and 12-fold per annum respectively compared with a 1.6-fold reduction in electronic media storage per year),[19] it is likely that in less than a decade, DNA-based storage will be the medium of choice for archives with a horizon of ≥ 50 y. The cost of maintenance and storage must also be considered; DNA based data storage, which requires negligible maintenance, presents a significant advantage in this context compared with the current gold standard of archival magnetic tape which requires maintenance and regular data transfers. Indeed, assuming that tape archives have to be read and rewritten every 5–10 y, current DNA based storage is cost-effective over a ~600–5,000-y horizon.

In a serendipitous coincidence, the Goldman et al.[4] study follows in the aftermath of a controversial year-long analysis and exposé on the unbridled energy consumption of data centers such

as Google, eBay and Facebook, published recently by the *New York Times* in an article entitled, "The Cloud Factories: Power, Pollution and the Internet."[20] In contrast, DNA mediated storage provides an eco-friendly archival data storage solution that begs the question whether future data storage solutions lie in cloud accessible bio-banks rather than energy hungry data centers.

However, DNA-based storage is itself not without limitations, including the lack of random access reads, as DNA sequencers read information sequentially; the "write-once" nature of DNA; and its latency, making it practical only for archival solutions. Indeed, a significant challenge facing long-term DNA-based storage is the ability to decode the data in the distant future. Egyptian hieroglyphics for example, widely believed to be the most ancient form of writing, dating back ~3300 BC, were decoded only as a result of the Rosetta stone, inscribed with the equivalent Greek text—without this ancient translation tool we would not be able to interpret the characters and symbols which constitute this ancient language. Therefore, without an equivalent molecular Rosetta stone, long-term archival data are likely to be completely unintelligible 5,000 thousand years from now (the time-frame for which current DNA-based data storage is cost effective). However, aside from this, which is after all a limitation inherent to all long-term archival strategies, many of the other more pressing concerns are, even now, beginning to be addressed. Random access, for example, might be facilitated if sequence fragments between barcodes are PCR amplified with a file allocation tube used as a file to barcode index mechanism. The challenge of rewritable DNA storage could be circumvented by utilizing PCR amplification to create multiple redundant backups. Furthermore, researchers at Stanford recently detailed a method for rewritable DNA[21] which uses bacteriophage enzymes called recombinases to flip a particular DNA segment back and forth to represent a binary 0 or 1. Although still in the early stages of development, the authors are currently scaling up to a byte and are reducing the latency involved (currently one hour for 1 bit of memory).

Therefore, despite the economic impracticality of DNA storage in 2013, this surprisingly simple idea has the potential to reshape the global face of data storage in the not too distant future (**Fig. 1**). Move over, Moore's law—make way for life's law.

Disclosure of Potential Conflicts of Interest

No potential conflict of interest was disclosed.

Acknowledgments

Both R.D.S. and A.O'D. are PIs on the FP7 Marie Curie ClouDx-i project. R.D.S. is an ESCMID Research Fellow.

References

1. Yang H. Genomics and the 'century of biology'. FEBS J 2012; 279:35.
2. Gantz J, Reinsel, D. The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. IDC iView: IDC Analyze the Future, 2012.
3. Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science 2012; 337:1628; PMID:22903519; http://dx.doi.org/10.1126/science.1226355.
4. Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 2013; 494:77-80; PMID:23354052; http://dx.doi.org/10.1038/nature11875.
5. Kelly TJ Jr., Smith HO. A restriction enzyme from Hemophilus influenzae. II. J Mol Biol 1970; 51:393-409; PMID:5312501; http://dx.doi.org/10.1016/0022-2836(70)90150-6.
6. Smith HO, Wilcox KW. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. J Mol Biol 1970; 51:379-91; PMID:5312500; http://dx.doi.org/10.1016/0022-2836(70)90149-X.
7. Little JW, Zimmerman SB, Oshinsky CK, Gellert M. Enzymatic joining of DNA strands, II. An enzyme-adenylate intermediate in the dpn-dependent DNA ligase reaction. Proc Natl Acad Sci U S A 1967; 58:2004-11; PMID:4295585; http://dx.doi.org/10.1073/pnas.58.5.2004.
8. Zimmerman SB, Little JW, Oshinsky CK, Gellert M. Enzymatic joining of DNA strands: a novel reaction of diphosphopyridine nucleotide. Proc Natl Acad Sci U S A 1967; 57:1841-8; PMID:4291949; http://dx.doi.org/10.1073/pnas.57.6.1841.
9. Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol 1987; 155:335-50; PMID:3431465; http://dx.doi.org/10.1016/0076-6879(87)55023-6.
10. Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, Brand TB, et al. Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science 2007; 317:111-4; PMID:17615355; http://dx.doi.org/10.1126/science.1141758.
11. Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, et al. Sequencing the nuclear genome of the extinct woolly mammoth. Nature 2008; 456:387-90; PMID:19020620; http://dx.doi.org/10.1038/nature07446.
12. Thomas MG. The flickering genes of the last mammoths. Mol Ecol 2012; 21:3379-81; PMID:22953331; http://dx.doi.org/10.1111/j.1365-294X.2012.05594.x.
13. Davis J. Microvenus + Biology, contemporary art, genetics. Art J 1996; 55:70-4; http://dx.doi.org/10.2307/777811.
14. Sleator RD. The story of Mycoplasma mycoides JCVI-syn1.0: the forty million dollar microbe. Bioeng Bugs 2010; 1:229-30; PMID:21327053; http://dx.doi.org/10.4161/bbug.1.4.12465.
15. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 2010; 329:52-6; PMID:20488990; http://dx.doi.org/10.1126/science.1190719.
16. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. Anal Chem 2011; 83:4327-41; PMID:21612267; http://dx.doi.org/10.1021/ac2010857.
17. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1953; 171:737-8; PMID:13054692; http://dx.doi.org/10.1038/171737a0.
18. Fehér T, Burland V, Pósfai G. In the fast lane: large-scale bacterial genome engineering. J Biotechnol 2012; 160:72-9; PMID:22406111; http://dx.doi.org/10.1016/j.jbiotec.2012.02.012.
19. Carr PA, Church GM. Genome engineering. Nat Biotechnol 2009; 27:1151-62; PMID:20010598; http://dx.doi.org/10.1038/nbt.1590.
20. Glanz J. The Cloud Factories: Power, Pollution and the Internet. In: Abramson J, ed. New York Times. New York: Arthur Ochs Sulzberger, Jr., 2012.
21. Bonnet J, Subsoontorn P, Endy D. Rewritable digital data storage in live cells via engineered control of recombination directionality. Proc Natl Acad Sci U S A 2012; 109:8884-9; PMID:22615351; http://dx.doi.org/10.1073/pnas.1202344109.