HHS Public Access

Author manuscript

IEEE Workshop Microelectron Electron Devices. Author manuscript; available in PMC 2019 September 20.

Published in final edited form as:

IEEE Workshop Microelectron Electron Devices. 2017 April ; 2017: . doi:10.1109/WMED.

2017 7916922.

A Coding Scheme for Nucleic Acid Memory (NAM)

Kelsey Suyehira,

Department of Computer Science, Boise State University, Boise, ID, USA,
kelseysuyehira@u.boisestate.edu

Simon Llewellyn,

Department of Computer Science, Boise State University, Boise, ID, USA, llewelsd@gmail.com

Reza M. Zadegan [Advisor],

Micron School of Materials Science & Engineering, Boise State University, Boise, ID, USA,
rezazadegan@boisestate.edu

William L. Hughes [Advisor],

Micron School of Materials Science, & Engineering, Boise State University, Boise, ID, USA,
willhughes@boisestate.edu

Tim Andersen [Advisor]

Department of Computer Science, Boise State University, Boise, ID, USA,
tandersen@boisestate.edu

Abstract

The global demand for digital data is projected to be greater than the supply of semiconductor grade silicon in 2040 [1]. When combined with the need to archive information [2], nucleic acids are being explored as an alternative memory material [1–7]. According to a recent study, the information density of nucleic acid memory (NAM) is 1000 times greater than flash memory and has the ability to last for hundreds of years [1]. Presented here is an algorithm for converting digital data into unique DNA sequences for glacial storage. Biologically inspired, our coding scheme maps hexadecimal characters to sequences of three DNA nucleotides. This mapping avoids repeating sequences and start codons, which could have adverse effects. We were able to encode and decode various file types without error.

Keywords

DNA; NAM; Nucleic Acid Memory; Data storage; Coding

I. Introduction

DNA encodes and stores the genetic information needed to create and maintain life. When used to store non-biological information, DNA is called nucleic acid memory (NAM). When compared to electronic and magnetic memory, NAM is extremely information dense and energy efficient [1]. Given the practical storage density of DNA, today's global storage needs ($\sim 10^{22}$ bits) could fit into a $10 \times 10 \times 10 \text{ cm}^3$ box [1]. Table 1 compares NAM to silicon-based memory materials, which is discussed in a prior study [1]. However, one

should keep in mind that the presented benchmarks usually reflect a conservative estimate on how DNA is stored inside living cells, and therefore do not directly translate to in-silico.

The quaternary code of DNA includes adenine (A), thymine (T), cytosine (C), and guanine (G). To store information in NAM, these nucleotides are sequentially synthesized to create information strings that have four letter alphabets. Once written, sequences can be read using standard DNA sequencing technologies (Fig. 1).

In anticipation of routine DNA synthesis, sequencing, and storage problems, select sequences have been omitted from our NAM algorithm. In consideration of omitted sequences, a biologically inspired coding scheme has been created.

II. Related Work

Early attempts to encode information into DNA mapped symbols (letters, numbers [8] or a subset of ASCII characters) to nucleotide triplets [9] or quadruplets [6]. The mappings of symbols were arbitrary and were only used to encode textual data. More recently, duplication codes [5] and Reed-Solomon codes [2] have been introduced for error correction. Random-access to the data via addressing schemes has also been introduced, which improves the read speed, increases the robustness of the data, and alleviates the need to recover all of the data in order to retrieve small sections [3, 7]. While these contributions are significant for digital DNA, the algorithms become considerably more complex. This complexity reflects a trade-off between additional features and information density. As the presented algorithm is biologically inspired, it is less complex than previous algorithms.

III. Methodology

A novel algorithm has been created for encoding and decoding digital DNA for nucleic acid memory (NAM). The algorithm was biologically inspired by protein translation and maps codons (i.e. triplets of three nucleotides) to hexadecimal characters. This mapping scheme was carefully constructed while considering specific sequences of nucleotides to avoid. Included in this list are special codes – known as start codons in biology – that signal initiation of biological processes. We have omitted these codons to ensure that if/when our data encoded DNA comes into contact with biological contents it would not trigger bio-activities. In excluding some codons, we have also shown that it is possible to avoid certain sequences that would lead to complications in read and write processes.

The results are that a codon option exists, to follow the next hexadecimal character to be encoded. This mapping scheme is shown in Table 2 along with the list of codons to avoid. During the encoding process, each codon is chosen from a list of options that avoid undesirable sequences. For example, if a codon created a sequence of three repeating nucleotides, the algorithm will choose again from the list of possible options. If no working codon could be identified, the system backtracks one hexadecimal character to pick a new option. The algorithm separated the data into 1-megabyte sections and then indexed each section, where the numerical index is encoded as is outlined above.

IV. Validation

The algorithm was tested by encoding various files including a jpg photo of a sunrise (3.4MB) and a randomly generated binary file (1GB). We also encoded/decoded three files tested by Goldman *et al.* including 154 of Shakespeare's sonnets (120KB), a pdf of Watson and Crick's paper (364KB) and an MP3 audio of Martin Luther King Jr.'s "I Have a Dream" speech (7.6MB) [5]. We saved the encoded data as sequences of the characters A, T, C, and G. From those sequences, the data was decoded back into its original form with 100% accuracy.

V. Conclusion

Nucleic acid memory (NAM) is attractive for archival storage of digital information. A stable and efficient coding scheme requires considering the issues of working with DNA as an organic and biological material. Our algorithm avoids creating sequences with repeating nucleotides or biologically active codes. Future work will adopt error-correction, as well as improve the addressing scheme.

VI. Challenges And Scalability Considerations

To become industrially practiced, alternative memory materials – e.g. suggested NAM – need to be scalable, and when compared with current technology provide improved properties and economical encouragement. Therefore such materials need to be suitable for high-volume manufacturing, have great information stability, limit the energy consumption, and reduce environmental impacts, while being economically scalable. Therefore, the memory community needs to carefully benchmark NAM against the current memory materials. A few aspects that are important for the industry include; NAM's (i) theoretical and practical write endurances, (ii) energy consumption, (iii) block size, (iv) random access, and (v) read and write limitations – including error rates, size of the information blocks, and speed. To address these issues, we encourage readers to review our published commentary article in Nature Materials [1]. Briefly explained, NAM provides information retention times that range from decades to millions of years, a volumetric density 1000 times greater, and an energy of operation 100 million times less than flash memory. Moreover, some attempts in the community have provided random access to NAM [7]. The major drawback of NAM as of today's technology could be the practical read and write latency. Although, the theoretical speed – given by the living cells – could be satisfactory, current chemical synthesis (write) and sequencing (read) of NAM are expensive, both time-wise and computationally. Therefore, further studies to improve read and write of NAM are recommended.

VII. Acknowledgments

This project was supported in part by NSF Grant No. CMMI-1344915 and the National Institute of General Medical Sciences of the National Institutes of Health under Award Number K25GM093233. This work was also supported in part by FAME, one of six centers of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

REFERENCES

- [1]. Zhirnov V, Zadegan RM, Sandhu GS, Church GM, & Hughes WL 2016 Nucleic acid memory. *Nature materials*, 15(4), 366–370 [PubMed: 27005909]
- [2]. Grass RN, Heckel R, Puddu M, Paunescu D, & Stark WJ 2015 Robust Chemical Preservation of Digital Information on DNA in Silica with ErrorCorrecting Codes. *Angewandte Chemie International Edition*, 54(8), 2552–2555. [PubMed: 25650567]
- [3]. Bornholt J, Lopez R, Carmean DM, Ceze L, Seelig G, & Strauss K 2016, March. A DNA-based archival storage system. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems* (pp. 637–649). ACM.
- [4]. Church GM, Gao Y, & Kosuri S 2012 Next-generation digital information storage in DNA. *Science*, 337(6102), 1628–1628. [PubMed: 22903519]
- [5]. Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, & Birney E 2013 Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 494(7435), 77–80. [PubMed: 23354052]
- [6]. Jiménez-Sánchez A 2013 A proposal for a DNA-based computer code. *International Inventions Journal Biochemistry and Bioinformatics*, 1, 1–4.
- [7]. Yazdi SHT, Yuan Y, Ma J, Zhao H, & Milenkovic O 2015 A rewritable, random-access DNA-based storage system. *Scientific reports*, 5, 14138. [PubMed: 26382652]
- [8]. Clelland CT, Risca V, & Bancroft C 1999 Hiding messages in DNA microdots. *Nature*, 399(6736), 533–534. [PubMed: 10376592]
- [9]. Wong PC, Wong KK, & Foote H 2003 Organic data memory using the DNA approach. *Communications of the ACM*, 46(1), 95–98.

Fig. 1. Digital data is encoded into DNA code. DNA sequences are then synthesized. Encoded information then decoded using standard sequencing technologies to retrieve the original data.

Table 1.

Comparison Of Data Storage Techniques [1].

	Read/Write Latency *	Retention *	Volumetric Density *
Hard disk	~3–5 ms per bit	>10 years	$\sim 10^{13}$ bit cm^{-3}
DRAM	<10 ns per bit	~64 ms	$\sim 10^{13}$ bit cm^{-3}
Flash Memory	~100 μs per bit	~10 years	$\sim 10^{16}$ bit cm^{-3}
DNA *	<100 μs per bit	>100 years	$\sim 10^{19}$ bit cm^{-3}

* Numbers present potential developed technology based on data acquired from in-vivo (in living cells) systems.

Table 2.

Hexadecimal To Codon Mapping Scheme

Hexadecimal	Codons	Hexadecimal	Codons
0	AAC, GAC	8	AGT, GGT
1	AAG, GAG	9	AGC, GGC, CCG
2	AGG, GTC, TCT	a	GAA, CGG
3	TCG, CGA	b	TAA, CAA
4	ACT, GCT, TCC	c	TAC, CCT, ATC
5	ACC, GCC, CGT	d	TAG, CGC
6	ACG, GCG	e	TTA, CTA, GTT
7	AGA, GGA	f	TTC, CTC, GTA
Codons to avoid	AAA, TTT, CCC, GGG, AAT, ATA, ATT, ATG, TAT, TTG, CAT, CAC, CAG, CTT, CTG, GTG		