International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

A Survey on Efficient Compression Technique for Generating DNA Sequences

S. Kavitha[1], Herold Lucia P[2]

[1, 2]Faculty of Computer Science, APL Global School, Chennai, India

Abstract: ***Demand for data storage is growing exponentially, but the capacity of existing storage media is not keeping up. Using DNA to archive data is an attractive possibility because it is extremely dense, with a raw limit of 1 Exabyte/mm3 (109 GB/mm3), and long-lasting, with observed half-life of over 500 years. This paper presents architecture for a DNA-based archival storage system. It is structured as a key-value store, and leverages common biochemical techniques to provide random access. We also propose a new encoding scheme that offers controllable redundancy, trading off reliability for density. We demonstrate feasibility, random access, and robustness of the proposed encoding with wet lab experiments involving 151 KB of synthesized DNA and a 42 KB random- access subset, and simulation experiments of larger sets calibrated to the wet lab experiments. Finally, we highlight trends in biotechnology that indicate the impending practicality of DNA storage for much larger datasets.***

Keywords: Data Storage, Data Compression, Big Data, Data Synthesis

1. Introduction

The Demand for Data storage is growing exponentially but the capacity of the existing storage media is not able to meet the demands. Each day huge amount of information is created, used and shared. We in the Era of Big Data, are generating a data flow of more than 8000 Exabyte's and will reach around 40000 Exabyte's in 2020. The large amount of data generated need large databases to store them, which arises to two major issues: Encoding of data and the time required to process them. The data being generated is always to have some redundancy in it and hence we need some compression methods to remove redundant data and then reduce the data size for storage.

The Data centers of various Giant Organizations are releasing large amount of Heat energy with regards to the storage of data which results as one of the major reason for Global Warming. Moreover Silicon and non-biodegradable materials used pollute the environment as they are toxic to humans.

DNA (Deoxyribonucleic acid), the element of all known living organisms tends to become an area of research as the future storage medium of digital data. DNA could be seen as a potential medium for storage purposes as the major functionality of the molecule is based on four sequential bases {A,C,G,T} similar to the 0's and 1's in a computer. The storage capacity in a DNA molecule is unbelievable due to the high density feature. It is stated that 4 grams of DNA is enough to store all of the world's data generated for one day. DNA, a biological molecule is environment friendly and hence it is not toxic to the society.

Researchers have encoded data into a DNA as sequences and successfully retrieved it without any errors. My research is to take up a dataset and develop compression algorithm to encode into DNA sequences which will not result in errors during mutations.

2. Background on DNA Manipulation

DNA basics. Naturally occurring DNA consists of four types of nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). A DNA strand, or *oligonucleotide*, is a linear sequence of these nucleotides. The two ends of a DNA strand, referred to as the 5′and 3′ ends, are chemically different. DNA sequences are conventionally represented starting with the 5′ nucleotide end. The interactions between different strands are predictable based on sequence. Two single strands can bind to each other and form a double helix if they are complementary: A in one strand aligns with T in the other, and likewise for C and G. The two strands in a double helix have opposite directionality (5′ end binds to the other strand's 3′ end), and thus the two sequences are the "reverse complement" of each other. Two strands do not need to be fully complementary to bind to one another. Such *partial* complementarity is useful for applications in DNA nanotechnology and other fields, but can also result in undesired "crosstalk" between sequences in complex reaction mixtures containing many sequences.

Selective DNA amplification with polymerase chain reaction (PCR). PCR is a method for exponentially amplifying the concentration of selected sequences of DNA within a pool. A PCR reaction requires four main components: the template, sequencing primers, a thermo stable polymerase and individual nucleotides that get incorporated into the DNA strand being amplified. The template is a single- or double-stranded molecule containing the (sub)sequence that will be amplified. The DNA sequencing primers are short synthetic strands that define the beginning and end of the region to be amplified. The polymerase is an enzyme that creates double-stranded DNA from a single-stranded template by "filling in" individual complementary nucleotides one by one, starting from a primer bound to that template. PCR happens in "cycles", each of which doubles the number of templates in a solution. The process can be repeated until the desired number of copies is created.

DNA synthesis. Arbitrary single-strand DNA sequences can be synthesized chemically, nucleotide by nucleotide.

The *coupling efficiency* of a synthesis process is the probability that a nucleotide binds to an existing partial strand at each step of the process. Although the coupling efficiency for each step can be higher than 99%, this small error still results in an exponential decrease of product yield with increasing length and limits the size of oligonucleotides that can be efficiently synthesized to about 200 nucleotides. In practice, synthesis of a given sequence uses a large number of parallel start sites and results in many truncated byproducts (the dominant error in DNA synthesis), in addition to many copies of the full length target sequence. Thus, despite errors in synthesizing any specific strand, a given synthesis batch will usually produce many perfect strands. Moreover, modern array synthesis technique can synthesize complex pools of nearly 105 different oligonucleotides in parallel.

DNA sequencing
There are several high-throughput sequencing techniques, but the most popular methods (such as that used by Illumina) use DNA polymerase enzymes and are commonly referred to as "sequencing by synthesis". The strand of interest serves as a template for the polymerase, which creates a complement of the strand. Importantly, *fluorescent* nucleotides are used during this synthesis process. Since each type of fluorescent nucleotide emits a different color, it is possible to read out the complement sequence optically. Sequencing is error-prone, but as with synthesis, in aggregate, sequencing typically produces enough precise reads of each strand.

3. Existing Encodings

Early work in DNA storage used encodings simpler than the one we describe above. For example, Bancroft et al. translate text to DNA by means of a simple ternary encoding: each of the 26 English characters and a space character maps to a sequence of three nucleotides drawn from A, C, and T (so exactly 33 = 27 characters can be represented). The authors successfully recovered a message of 106 characters, but this encoding suffers substantial overheads and poor reliability for longer messages.

Goldman encoding
This encoding splits the input DNA nucleotides into overlapping segments to provide fourfold redundancy for each segment. Each window of four segments corresponds to a strand in the output encoding. The authors used this encoding to successfully recover a 739 kB message. We use this encoding as a baseline because it is, to our knowledge, the most successful published DNA storage technique. In addition, it offers a tunable level of redundancy, by reducing the width of the segments and therefore repeating them more often.

This encoding incorporates redundancy by taking the exclusive-or of two payloads to form a third. Recovering any two of the three strands is sufficient to recover the third. The strands of the same length.

XOR Encoding
While the Goldman encoding provides high reliability, it also incurs significant overhead: each block in the input string is repeated four times. We propose a simple new encoding that provides similar levels of redundancy to prior work, but with reduced overhead.

Our encoding, shown in provides redundancy by a simple exclusive-or operation at the strand level. We take the exclusive-or $A \oplus B$ of the payloads A and B of two strands, which produces a new payload and so a new DNA strand. The address block of the new strand encodes the addresses of the input strands that were the inputs to the exclusive-or; the high bit of the address is used to indicate whether a strand is an original payload or an exclusive-orstrand. This encoding provides its redundancy in a similar fashion to RAID 5: any two of the three strands A, B, and $A \oplus B$ are sufficient to recover the third.

The reliability of this encoding is similar to that of Gold- man. We show that we successfully recovered objects from both encodings in a wet lab experiment. How- ever, the theoretical density of this encoding is much higher than Goldman—where in their encoding each nucleotide re- peats (up to) four times, in ours each nucleotide repeats an average of 1.5 times. In practice, the density difference is lower, due to the overheads of addressing and primers that are constant between the two encodings. Simulation results show our encoding to be twice as dense as that of Goldman, and for all practical DNA synthesis and sequencing technologies, it provides equivalent levels of reliability.

Tunable Redundancy
Recent work in approximate storage shows that many applications do not need high-precision storage for every data structure. For example, while the header data of a JPEG file is critical to successful decoding, small errors in the payload are tolerable at the cost of some decoding imprecision.

One key advantage of our encoding is that the level of redundancy is tunable at a per-block granularity. For critical data, we can provide high redundancy by pairing critical

File Recovery was successfully recovered all four files from the sequenced DNA. Three of the files were recovered without manual intervention. One file – cat.jpg encoded with the Goldman encoder – incurred a one-byte error in the JPEG header, which we fixed by hand. As described, the design of the Goldman encoder provides no redundancy for the first and last bytes of a file, and so this error was due to random substitution in either sequencing or synthesis. We could mitigate this error scenario by trivially extending that algorithm to wrap the redundant strands past the end of the file and back to the beginning.

4. Conclusion

DNA-based storage has the potential to be the ultimate archival storage solution: it is extremely dense and durable.
While this is not practical yet due to the current state of DNA synthesis and sequencing, both technologies are improving at an exponential rate with advances in the biotechnology in- dustry. Given the impending limits of silicon technology, we believe that hybrid silicon and

biochemical systems are worth serious consideration: time is ripe for computer architects to consider incorporating biomolecules as an integral part of computer design. DNA-based storage is one clear example of this direction. Biotechnology has benefited tremendously from progress in silicon technology developed by the com- puter industry; perhaps now is the time for the computer industry to borrow back from the biotechnology industry to advance the state of the art in computer systems.

References

[1] L. Adleman. Molecular computation of solutions to combina- torial problems. *Science*, 266(5187):1021–1024, 1994.

[2] M. E. Allentoft, M. Collins, D. Harker, J. Haile, C. L. Oskam, M.L. Hale, P. F. Campos, J. A. Samaniego, M. T. P. Gilbert, E.Willerslev, G. Zhang, R. P. Scofield, R. N. Holdaway, and M. Bunce. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1748):4724– 4733, 2012.

[3] C. Bancroft, T. Bowler, B. Bloom, and C. T. Clelland. Long- term storage of information in DNA. *Science*, 293(5536): 1763–1765, 2001.

[4] R. Carlson. Time for new DNA synthesis and sequencing cost curves. http://www.synthesis.cc/2014/02/time-for-new-cost-curves-2014.html, 2014.

[5] Y.-J. Chen, N. Dalchau, N. Srinivas, A. Phillips, L. Cardelli, D.Soloveichik, and G. Seelig. Programmable chemical con- trollers made from DNA. *Nature Nanotechnology*, 8(10):755– 762, 2013.

[6] G. M. Church, Y. Gao, and S. Kosuri. Next-generation digital information storage in DNA. *Science*, 337(6102):1628, 2012.

[7] C. T. Clelland, V. Risca, and C. Bancroft. Hiding messages in DNA microdots. *Nature*, 399:533–534, 1999.

[8] ExtremeTech. New optical laser can increase DVD storage up to one petabyte. http://www.extremetech.com/computing/159245-new-optical-laser-can-increase-dvd-storage-up-to-one-petabyte, 2013.

[9] D. G. Gibson, J. I. Glass, C. Lartigue, V. N. Noskov, R.-Y.Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M.M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N.Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L.Young, Z.-Q. Qi, T. H. Segall-Shapiro, C. H. Calvey, P. P. Parmar, C. A. Hutchison, H. O. Smith, and J. C. Venter. Cre- ation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 329(5987):52–56, 2010.

[10] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B.Sipos, and E. Birney. Towards practical,high-capacity, low- maintenance information storage in synthesized DNA. *Nature*, 494:77–80, 2013.

[11] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem. Int. Ed.*, 54:2552–2555, 2015.

[12] Q. Guo, K. Strauss, L. Ceze, and H. Malvar.High-density image storage using approximate memory cells. In *ASPLOS*, 2016.

[13] D. Huffman. A method for the construction of minimum- redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[14] IDC. Where in the world is storage. http://www.idc.com/downloads/where_is_storage_ infographic_243338.pdf, 2013.

[15] S. Kosuri and G. M. Church. Large-scale de novo DNA synthesis: technologies and applications. *Nature Methods*, 11:499–507, 2014.

[16] A. Leier, C. Richter, W. Banzhaf, and H. Rauhe. Cryptography with DNA binary strands. *Biosystems*, 57(1):13–22, 2000.

[17] M. D. Matteucci and M. H. Caruthers. Synthesis of deoxy- oligonucleotides on a polymer support. *Journal of the Ameri- can Chemical Society*, 103(11):3185–3191, 1981.

[18] R. Miller. Facebook builds exabyte data centers for cold storage. http://www.datacenterknowledge.com/archives/2013/01/18/facebook-builds-new-data-centers-for-cold-storage/, 2013.

[19] R. A. Muscat, K. Strauss, L. Ceze, and G. Seelig. DNA-basedmolecular architecture with spatially localized components. in *International Symposium on Computer Architecture, 2013*.

[20] T. P. Niedringhaus, D. Milanova, M. B. Kerby, M. P. Snyder, and A. E. Barron. Landscape of next-generation sequencing technologies. *Anal. Chem.*, 83:4327–4341, 2011.