EasyChair Preprint

№ 1885

A Survey Paper on DNA-Based Data Storage

Shubham Taluja, Jagrit Bhupal and Siva Rama Krishnan S

November 8, 2019

A Survey Paper on DNA-Based Data Storage

Shubham[1], Jagrit[2] Siva Rama Krishnan S[3]
[1,2]B.Tech I.T,
[3]Assistant Professor (Senior)
[1, 2, 3]School of Information Technology and Engineering
VIT, Vellore

Abstract- On average, 2.5 quintillion bytes of data are generated on the internet every day. So, the demand for data storage is growing exponentially, but the capacity of existing storage media is not keeping up. A revolution in the field of data storage is the need of an hour. This paper surveys a very unique technique of storing digital data in DNA sequences. The idea is to replicate nature's way of storing information. It has been around with us for ages. This paper elaborates on the procedure, applications, and challenges that are associated with this fictitious idea of data storage. 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 an observed half-life of over 500 years.

Keywords- DNA Storage, Digital Storage, Random Access, High Capacity Data storage, Archival Data, long-term data storage

I. OBJECTIVE

This paper reviews almost every milestone related to the field of DNA storage. From the pros and cons to how and why to everything possible to everything which hinders the growth of such technology. This paper also discusses the fact of how much data storage is possible at the current stage of time apart from what theoretical concepts claim to do.

II. INTRODUCTION

DNA has been around us from the beginning and will remain till eternity. DNA has some very remarkable properties like high capacity storage and amazing longevity. The amount of information that could be stored even in a single gram of DNA is just enormous, precisely 215 million GB in a single gram. The number of DNA cells that exists inside a single human can store much more data than that the whole of humanity has generated so far. There are different studies on how long this digital data can remain preserved within the genome, though these studies use kinetically accelerated methods, the results have successfully retrieved data with high accuracy from about 2000-4000 years of artificially aged data.

III. METHODOLOGY

Let's look into the procedure of storing a bit (1 or 0) of information in a DNA molecule. Well, the process is fairly simple but to understand the process one must understand the structure of DNA. It is a molecule with a double-helical structure consisting of four bases A, C, G, T (adenine, cytosine, guanine or thymine). A-T and C-G are often termed as the base pairs as they form a bond together. The synthetic form of the genome is then sequenced in the vitro (in glass) form using encoding schemes such as naming one bond pair as 0 and other as 1to store digital data. The other possible encoding technique is by naming all the bases as a different binary string (A-00, T-01, C-10, and G-11). Different encoding schemes have their advantages and disadvantages. So they are used according to the type of data being stored or to be retrieved.

IV. LITERATURE SURVEY

Jonathan [1] discussed how DNA is an excellent method for data storage and also addresses two major problems associated with it. The first retrieval of data from the genome is a very tedious process although that can be expected to improve. Another being the cost factor since such technology could be very attracting and thus cost goes up.

James Bornholt et al [2] examined the exponentially increasing demand for data storage and the unsatisfying capacity of existing storage media. To keep up with the demand, using DNA to archive data is an attractive possibility. In this paper, they presented an architecture for a DNA-based archival storage system which is designed as a key-value store leveraging common biochemical techniques to provide random access.

Yeongjae et al [3] describe the analogy between the digital data and genetic data. To start with digital information is stored as binary digits (bits i.e 0 and 1) and the genetic data is stored in the form of molecular polymers. These polymers consist of four bases (A, C, T, G) each pair corresponding to a bit of information.

S.M. Hossein et al [4] described the first DNA-based storage architecture that enables random access to data blocks and rewriting of information stored at arbitrary locations within the blocks. Their system is based on new constrained coding techniques with DNA editing methods that ensure data reliability, specificity, and sensitivity of access, and at the same time provide exceptionally high data storage capacity.

Nick et al [5] explained the maintenance of DNA based storage which requires no maintenance other than a cold dark and dry environment which is true for any biological or chemical compound. Some other properties were also discussed such as making copies is highly efficient which makes it excellent for backups and transportation. Thus DNA storage can be termed as a highly potential and practical and cost-effective solution for archiving data which is rarely accessed since the retrieval speed is low.

Meinolf Blawat et al [6] proposed a forward error correction scheme that can cope with all error types of today's DNA synthesis, amplification and sequencing processes, e.g. insertion, deletion and swap errors. After performing successful experiments, they were able to store and retrieve error-free 22 Mbyte of digital data in synthetic DNA.

George et al [7] developed a novel encoding technique for synthesizing DNA. The results obtained had a lot of advantages when compared to previous traditional storage methods. The baseline of this unique technique was to encode 1 bit per base instead of storing two, this allowed to encode messages with more possibilities so that difficult sequences can be avoided.

S. M. Hossein et al [8] proposed DNA-based data storage as an emerging non-volatile memory technology of potentially unmatched durability, density, and replication efficiency. As for now, existing architectures of DNA Storage enables only reading and writing but no random-access and error-free data recovery. In this paper, they implemented a portable, random-access platform in practice using nanowire sequencers.

Boris et al [9] discussed the biology encoding methodology of digital storage in DNA and also proposed novel techniques about how DNA storage can be utilized to hide data. It discusses how DNA can be used to store anything of value due to its remarkable properties one of which includes non-tampering watermark which provides a basis for distinguishing the original owner of the information. Further, they also address methods on strengthening the watermark and also show analysis of results.

C. Mayer et al [10] proposed that biopolymers can be an attractive alternative to store and circulate information. By using differential kinetics of hydraulic deamination reactions of cytosine and its naturally occurring derivate, they demonstrate how multiple layers of data can be stored in a single DNA template.

Robert et al [11] presented proof that the digital data stored in DNA can be retrieved after a very long time. They experimented by storing information in DNA segments encapsulated with error-correcting codes, then treating the mixture to some very harsh conditions such as high temperature and by accelerating the aging process. They monitored the results for kinetic decay over time and then successfully obtained the original information. They claim of artificial aging is equivalent to 2000 years in central Europe.

M. Fritz et al [12] stated that the data storage costs have become a large proportion of total cost in the creation and analysis of DNA sequence data. In this paper, they presented a new compression technique that efficiently compresses DNA sequences for storage.

Sharon et al [13] proposed how to store data on DNA spots on a glass plate. They experimented to showcase the possibility of their proposed technique. As a result, they successfully stored and retrieved about a terabyte of information on a single spot of DNA

Andy Extance [14] discussed how they can afford to store the genome sequences and other data the world was creating a very fast rate. They mentioned that DNA storage would be slow and it would take hours to store data by synthesizing DNA strings with a particular pattern of bases. And also if one wants to recover that information, it will take more hours as it will require a sequencing machine.

Siddaramappa et al [15] discussed the structure and the advancements in the field of genetic engineering and argues why data security is an important aspect of any storage system. They have introduces their cryptographic algorithm (XOR-based) which uses genes as the keys, the paper also mentions on how to overcome various cryptanalytic attacks and key length problems.

R Heckel et al [16] described how there has been a great demand for data storage as there have been different challenges to the current data centers and storage techniques. Due to this, there is a need to invent and implement new storage technologies beyond hard disks and memory chips. In this context, DNA is a suitable medium for archival data storage due to its longevity and vast information density. In this paper, they explained fundamentals about DNA Storage.

Luis et al [17] provided an overview of the whole procedure of the DNA as a data storage technology and the challenges that are being faced to procure the rise of mainstream acceptance. The uniqueness of this paper is that they have surveyed both the data storage in the DNA of the living cells as well as vitro DNA data storage.

Y. Erlich et al [18] proposed that DNA can provide large-capacity information storage. They stated how the present techniques have not been able to implement this concept. They proposed a technique named DNA Fountain which can approach the conceptual maximum for information stored per nucleotide.

Lee et al [19] conducted experiments using previously known encoding schemes and concludes that reliable data retrieval is possible even if only 10 copies were generated initially per sequence, which in turn increases the density. Although this paper does not discuss anything about accessing information from that dense and complex pool.

M. Arita [20] stated that it is a high time when one must try writing data or information into DNA as it is suitable for large-capacity information storage. In this paper, he surveys the techniques for designing codewords using DNA.

V. CHALLENGES

The main obstacles which are faced while storing data in DNA are cost and efficiency. The process of encoding data into DNA is incredibly slow. The rate of encoding data is about 400 bytes per second. This is millions of times slower than the microsecond timescales for reading and writing bits in a silicon memory chip. As estimated by Microsoft, a data storage technology to become feasible in practice must encode the information at a rate of 100 MB per second. The cost required for synthesizing DNA molecules is very large.

The other challenges with DNA storage are, how one will identify the DNA strand in which the file they are looking for is contained. After identifying those strands, how to remove these strands so that the files inside them can be read. These operations must be done without destroying the strands. To overcome these challenges, researchers [21] introduced two techniques named as DNA Enrichment and Nested Separation or together called as Dense.

The file identification task is tackled by using two, nested primer-binding sequences in which the system will first identify all the strands containing the initial binder sequence and then it will conduct a second search to find out the strands which contain second binder sequence. But this technique leads to an increase in the number of estimated file names from 30,000 to approximately 900 million immediately.

To extract the identified file, technique is to make lots of copies of the relevant DNA strands and then the entire sample is sequenced. Due to a large number of copies of the relevant DNA strands, their signal submerges the rest of the strands which makes it possible to identify the targeted DNA sequence. But these techniques are not efficient and fail to work if one is trying to retrieve data from a high-capacity database like DNA.

VI. APPLICATIONS

Startups around the world are making their way towards this new growing technology. A startup Catalog has reportedly announced that they have stored the whole Wikipedia's textual data of 16Gbs inside a tiny genome. The main application of DNA storage technology is to store archival data due to its two remarkable properties (i.e. high capacity and its stability for longer periods). Since the cost of synthesizing DNA at this stage of time is very high. The applications of archival data are restricted for only storing medical records, intelligence information or other legal or highly confidential data.

Since the DNA storage is offline storage and in no way connected with the internet, it is pretty safe against any cyber-attack. Cryptographic algorithms are also used to encrypt the data while storing based on the value of data. Some of the applications have evolved due to its secure nature, another startup Carverr is providing a service of storing bitcoin keys and passwords in DNA test tubes.

DNA storage could be the densest house of information even if data is stored at half of which is theoretically possible. Although researchers have proved by storing and retrieving data by achieving 85% of the theoretical limit. Searching and accessing data can be highly complex, a company has designed a molecular probe that can access information stored anywhere in the long sequences thus can be analogized with RAM or other random access high-speed memory devices. Further the synthesizing and the whole development process is eco-friendly which makes it more promising for upcoming future generation technology and maintaining sustainability.

VII. CONCLUSION

DNA storage is the densest and highest capacity data store and can sustain for the eternity if stored in the favorable conditions. Although at this stage of time the cost and efficiency drag the technology to achieving its true potential, by observing the growth of research in this field in terms of the data that was stored a few years back and its cost as compared to the ratio by this time. Also, due to its several advantages and its need as the digital universe is getting almost double every year. So, we can safely conclude that the technology will evolve in the upcoming time in terms of reduction of cost and achieving more availability.

VIII. REFERENCES

[1] Cox, J. P. (2001). Long-term data storage in DNA. *TRENDS in Biotechnology*, *19*(7), 247-250.

[2] Bornholt, J., Lopez, R., Carmean, D. M., Ceze, L., Seelig, G., & Strauss, K. (2016). A DNA-based archival storage system. *ACM SIGARCH Computer Architecture News*, *44*(2), 637-649.

[3] Choi, Y., Ryu, T., Lee, A. C., Choi, H., Lee, H., Park, J., ... & Kwon, S. (2019). High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Scientific reports*, *9*(1), 6582.

[4] Yazdi, S. H. T., Yuan, Y., Ma, J., Zhao, H., & Milenkovic, O. (2015). A rewritable, random-access DNA-based storage system. *Scientific reports*, *5*, 14138.

[5] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., & Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, *494*(7435), 77.

[6] Blawat, M., Gaedke, K., Huetter, I., Chen, X. M., Turczyk, B., Inverso, S., ... & Church, G. M. (2016). Forward error correction for DNA data storage. *Procedia Computer Science*, *80*, 1011-1022.

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

[8] Yazdi, S. H. T., Gabrys, R., & Milenkovic, O. (2017). Portable and error-free DNA-based data storage. *Scientific reports*, *7*(1), 5011.

[9] Shimanovsky, B., Feng, J., & Potkonjak, M. (2002, October). Hiding data in DNA. In *International Workshop on Information Hiding* (pp. 373-386). Springer, Berlin, Heidelberg.

[10] Mayer, C., McInroy, G. R., Murat, P., Van Delft, P., & Balasubramanian, S. (2016). An Epigenetics-Inspired DNA-Based Data Storage System. *Angewandte Chemie International Edition*, *55*(37), 11144-11148.

[11] Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., & Stark, W. J. (2015). Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angewandte Chemie International Edition, 54(8), 2552-2555.

[12] Fritz, M. H. Y., Leinonen, R., Cochrane, G., & Birney, E. (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, *21*(5), 734-740.

[13] Newman, S., Stephenson, A. P., Willsey, M., Nguyen, B. H., Takahashi, C. N., Strauss, K., & Ceze, L. (2019). High-density DNA data storage library via dehydration with digital microfluidic retrieval. *Nature communications*, *10*(1), 1706.

[14] Extance, A. (2016). How DNA could store all the world's data. *Nature News*, *537*(7618), 22.

[15] Siddaramappa, V., & Ramesh, K. B. (2019). DNA-Based XOR operation (DNAX) for data security using DNA as a storage medium. An Integrated Intelligent Computing, Communication and Security (pp. 343-351). Springer, Singapore.

[16] Heckel, R., Shomorony, I., Ramchandran, K., & David, N. C. (2017, June). Fundamental limits of DNA storage systems. In *2017 IEEE International Symposium on Information Theory (ISIT)* (pp. 3130-3134). IEEE.

[17] Ceze, L., Nivala, J., & Strauss, K. (2019). Molecular digital data storage using DNA. Nature Reviews Genetics, 1.

[18] Erlich, Y., & Zielinski, D. (2017). DNA Fountain enables a robust and efficient storage architecture. *Science*, *355*(6328), 950-954.

[19] Organick, L., Chen, Y. J., Ang, S. D., Lopez, R., Strauss, K., & Ceze, L. (2019). Experimental Assessment of PCR Specificity and Copy Number for Reliable Data Retrieval in DNA Storage. *bioRxiv*, 565150.

[20] Arita, M. (2003). Writing information into DNA. In *Aspects of Molecular Computing* (pp. 23-35). Springer, Berlin, Heidelberg.

[21] North Carolina State University. (2019, June 3). Key obstacles to scaling up DNA data storage. ScienceDaily. Retrieved October 22, 2019 from www.sciencedaily.com/releases/2019/06/190603124605.htm