INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of Second International Conference on Science and Technology (ICOST 2021)

DNA; Digital Data Storage Device

Kavia V
[1]*Student, College of Veterinary and Animal Sciences, Trissur, Kerala, India*
kavyasathyanathan@gmail.com

Abstract

The growth of various types of information happens day by day. The increased production of data, in turn, led to the invention of various data storage tools. The idea of storing data in DNA is now one of the major focuses. There are so many methods for storing data in DNA, the method put forward by Nick Goldman and Evan Birney has so many advantages over the other methods. They had converted data in the binary form into ternary by using improved Huffman coding of base 3. This helped to reduce the error rate that had produced in the earlier works.

Keywords: Increased data production,Data storage tools, Improved Huffman coding,Error rate

1. Introduction

There is a huge increase in the amount of data day by day. Our storage necessity increases by 50% every year. Data produced from various sources like important documents, data from social media all are increasing in a few times. To store the increasing amount of digital data, we should find the appropriate storage mechanisms. From the old days, we had used bones, rocks, and paper to store data. The evolution progressed after that through punched cards, magnetic tapes, floppy discs, gramophones and reached to CD'S, blue ray discs, DVD'S, and flash drives. These devices can easily get damaged by temperature, moisture and magnetic variations in the surrounding. All the above-mentioned devices are strong and non-biodegradable, it will lead to environment pollution. And also, it will loss efficacy with time. This led the scientists to discover more efficient DNA storage model.

1.1 Advantages from other digital data storing devices

An unbelievable amount of enormous data can be stored on DNA. The long shelf life and robustness are the properties of data. Even after thousands of years the information stored on DNA can be recovered by providing optimum conditions for keeping DNA in dry, dark and cold conditions [7]. Current data storing devices are more prone to data loss. The power usage by other data storing devices is very high, while for DNA, it is much more efficient than these devices. DNA is the genetic material of all living beings and is made up of polynucleotides, it contains a pentose sugar, a phosphate group and nitrogenous bases like adenine (A), guanine (G), cytosine(S) and thiamine (T). DNA has a double helical structure, the two separate strands are connected together by hydrogen bonds between purine and pyrimidine, and also by some base stacking interactions. Here 'A' invariably pairs with 'T', and 'G' invariably pairs with 'C. It is possible to store the world's whole data in a few grams of DNA. [5]. It is possible to get many copies of DNA by the usage of PCR (polymerase chain reaction). The durability and the long-term storage option making this to retain information for centuries. A huge amount of data can be stored in a very small space due to the high density of DNA. As in practical, the data is kept in long virtual DNA molecule but the encoding is done using short oligonucleotide strands, synthetically prepared, short strands allow to easily manipulate the data. Since data can store in DNA, it is secure as invisible to the human eye, and also, it can withstand extreme environmental conditions [8]. There are so many problems also arise with this

idea- storing data in DNA- like the cost of synthesis of oligonucleotide strands, and the inability to synthesise long DNA strands. Also, there may be errors in DNA synthetic machines during the synthesize of oligonucleotides.

2. Related Works

The possibility of DNA data storage was executed by Clell, Risca and Bancroft in 1999. They encoded data in short strands of DNA called microdots [2]. This technology had used in World War 2 to communicate data. The next research was published by Wong et al in 2003. They made small oligonucleotides in a manner similar to the arrangement of 1s and 0s in the ASCII code of computer language code. And also, they encoded data in *E.coi* cells. In 2011, research led by Church and Kosuri led to a more convenient method of DNA data storage. They assigned 0s as A or C, and 1s as T or G. But it made homopolymers throughout their oligonucleotide chain that made problems during DNA synthesising and sequencing.[3]. The method used by Church had several advantages over the past proposed models. They used purely in vitro approach that avoids cloning and stability issues in in vivo approach [6]. Later in 2013, work led by Goldman and E. Bierney created small pieces of oligonucleotide strands carrying image in jpg format and a video clip in mp3 format. This is an almost error-free way of encoding data in DNA. The conducted by Zhong and team had used base 4-digit method [10].

3. Encoding Strategy

The encoding used in DNA nucleotides breaks them into segments. Breaking them into segments provided a redundancy up to four folds. The data can be stored in DNA by mainly five steps.

nthesis

III.Storage

IV.Retrieval

V.Decoding

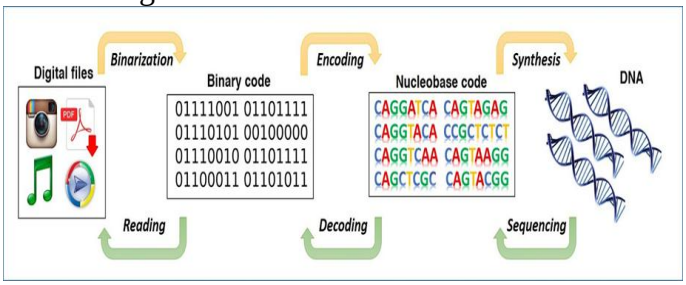

Fig.1. Describes the steps involved in the conversion of data from its original form to the encoded form

The data, present in the binary form, first converted to the nucleotide code, the corresponding nucleotides are synthesized by using DNA synthetic machines, and it can be stored for long time, data can be retrieved by sequencing the DNA and the sequenced material can be decoded in order to regain the data stored on DNA [8].

4. The Proposed System

There are so many proposed systems for storing data in DNA. One of the most prominent and the latest system is the method proposed by Goldman and et.al. They converted data, in the binary form, into a ternary format using specialized encoding format called 'the improved Huffman code'. This 'improved Huffman code' is coding the method derived from the old 'Huffman code' method, which converts data in the binary form into more compressed binary format (one method used for compression of data in the binary format in order to reduce the storage space). But in improved Huffman method, it converts data into a ternary format instead of binary format. Binary letters contain only the 1s and 2s, but in ternary form, the code is a three-number format, it contains 1s, 2s, and 0s. Goldman used this version of Huffman code and converted data in the binary format into a ternary format, the code containing only 1s, 2s, and 0s, for that he used a Huffman tree. Data compression is achieved by using a Huffman tree for encoding. No one can decode it without the original tree; this makes the data more secure. A lot of specialized equipment is needed for sequencing of DNA strand. In order to reduce data loss, there are two copies of data provided; in case of any data loss the other copy can be the alternate. Flexibility is considered as an advantage of this method; data manipulation can be done by the researcher to meet his criteria [4]. As we know, all the data can be converted to ASCII (American Standard Code for Information Interchange) format. ASCII format is nothing but a universal code used to store data in the digital format. Only data in the binary form can be transferred via digital means. ASCII code is an eight-letter binary code, which uses only 1s and 0s, and there are about 256 characters (standard) having ASCII code for each of them. Since there are eight digits in an ASCII code for a particular character, it reduces the storage capacity of a file, and it takes more space, time and energy to transfer data across the internet or other digital means. File

capacity is expressed in the bit, byte, kilobyte, and megabyte form. One byte equals eight bits. One ASCII code has eight characters; hence it is a one-byte character. Researchers think about the need for a binary code which takes less space and transfers a huge amount of data. This leads to the evolution of so many coding schemes, one of which is the code formed by Huffman. Huffman had used a tree (Huffman tree) to generate the code by using the frequency of each character in the data. Goldman had used an improved version of Huffman tree to translate the data into codes, but it is not for digital purposes, he converted data into a ternary format using improved Huffman coding scheme and then into a base - 3 for the format of nucleotides, and by using that he synthesized a new single-stranded DNA. The single-stranded DNA didn't contain any nucleotide repeats since it might confuse the DNA synthetic machines [1].

Goldman created a new method strategy (in the year 2013) by modifying Church's one bit per base system with the help of the improved base 3 Huffman coding'' (trits 0, 1 and 2). Here, he generated a triplet DNA code from the original binary code (0, 1) through a ternary code (0, 1, 2), all the four steps involved in this method is shown in the figure. Usually, binary digits have respective ASCII codes, when we convert binary to base 3 Huffman code that replaces each byte with base-3 digits (five or six). In order to avoid error created (homopolymer formation) during synthesis of DNA, each trit used here encoded with one of the three nucleotides different from the previous one used.

4.1 Encoding

Encoding is done by forming a data character's frequency table. New nodes are created for encoding the non-repeating nucleotides of the Huffman tree. These nodes will have three sub nodes (or children) as well as the incoming weight of each generation's parent is determined by the children's branches weight. For example, C is considered as the leftmost child, G in middle, and T as the rightmost if the weight of the incoming branch of a parent is A. like this, G is considered as the leftmost child, T in middle, and A as the rightmost if the weight of the incoming branch of a parent is C. The whole information split into overlapping segments of 100 nucleotides with an offset of 50 nucleotides from previous and create pairs of segments starting from the first segment [8].

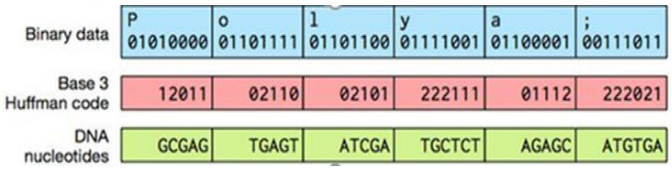

Fig.2.Converting binary data to DNA nucleotide

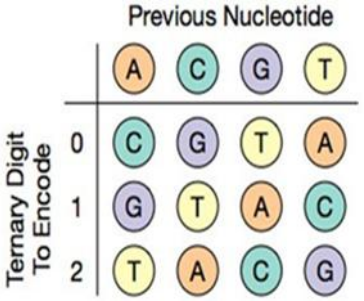

Fig.3. A rotating encoding to nucleotides avoids homopolymers, which will create errors

DNA strand is divided into chunks each of length 117 base pair (bp). 75 bases for each DNA information chunks were overlapped with four-fold redundancy to recover the data loss that occurred during synthesis and sequencing DNA. For the data security each redundant chunk is converted to reverse complement of the strand in every alternate chunks. Each DNA chunk is appended with data address blocks of 117 bases to determine the location of segment in overall data. One parity check bit is added for the intra file location and error detection. Total 153,335 DNA strings were generated. 33 nucleotide base primer was added to facilitate synthesis process and amplification. For details reader is referred to [Goldman et al. 2013]. As proof of concept, they used four different file types (739 kilobytes file size) and achieved 2.2 PB/g DNA storage density.

4.2 Decoding

The decoding the process is the simple reversal of the encoding. Whether the DNA belongs to the 1st or 2nd segment of the pair or whether the data is reverse complemented or not, can be identified from the first nucleotide. Remove the 1st nucleotide if it is A and the next four nucleotides will tell us about the segment number, and the next 100 nucleotides will be the data. For the confirmation of the type of segment the last nucleotide can be used. If the 1st nucleotide is C, then reverses the whole complement and then remove the first nucleotide, next four nucleotides will tell us about the segment

number, the data will be the next 100 nucleotides and the last a nucleotide can be used for confirmation of the type of segment. If G represents the first nucleotide, then first reverse the whole segment and remove first nucleotide, the next four nucleotides will tell us about the segment number. Then reverse complements next 100 nucleotides, these 100 nucleotides will be the data and the last nucleotide defines the segment's confirmation. If T represents the first nucleotide, then remove the 1st nucleotide, next four nucleotides will tell us about the segment number. Now, reverse complements next 100 nucleotides, these 100 nucleotides now will be the data, and the last nucleotide again defines the segment's confirmation. The 'TTTT' sequence will define the end of each file and after that new character will start from next nucleotide. Now the data can be converted back into the original file (Characters) by using the same Huffman tree. A single Huffman can be used for the whole data. To code different files different Huffman trees can also be used [8]. Now the DNA strands split into chunks. Each of the chunk has length approximately 117 base pairs as well as 75 bases in each DNA information chunks are overlapped with four-fold redundancy. This helps to recover data loss that will occur during synthesis and sequencing of DNA. In every alternate chunk, each redundant chunk gets converted to reverse complement of the strand, to secure data. Next, the determination of location of segments in overall data, each chunk was appended with data address blocks of 117 bases, for intra file location and error detection one parity check bit was added. Primer was added to facilitate synthesis process and amplification. Goldman & co-workers used four different file types; Shakespeare's sonnet in ASCII format, Watson & Crick's DNA double helical structure 1953 paper, a medium resolution colour photograph, a part of the audio of Martin Luther King's speech, which is in mp3 format; and they achieved 2.2 PB/g DNA storage density [9]. This will compress the whole data. The decoding process requires the original tree, without it no one can decode the data [8].

Conclusion

DNA is extremely dense and durable and thus it has the potential to be the ultimate archival storage form. By using DNA as a storage medium, storage of information in a very less space for the long-term purpose is possible. Similar to all revolutions in technology, it also has to face challenges in order to know its full potential. Since DNA has higher density, robustness, stability, and energy efficiency, it is the best archival device so far known. A minute amount of DNA can store all the information produced in the world, even though several breakthroughs will be required before it comes to the commercial sector. Future work could include compression schemes, parity checking, correcting errors in order to enhance density and safety.

References

Journals

[1] Ailenberg, M. and Rotstein, O.D., "An improved Huffman coding method for archiving text, images, and music characters in DNA",BioTechniques, vol. 47, no. 3, pp. 747‑754, 2009.

[2] Akram, F., Haq, I. U. and Laghari, A. T., "Trenda to store digital data in DNA: an overview", Springer Nature, 2018.

[3] Badlani, R. and Shrivastava, S., "Data storage in dna," International Journal of Electrical Energy, vol. 2, no. 2, 2014.

[4] Bertone, P., Birney, E., Chen, S., Dessimoz, C., Goldman, N., LeProust, E. M. and Sipos, B., "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA", Nature, 494(7435), 77-80, 2013.

[5] Chaczko, Z., Hakami, A.A. and Kale, A., "Review of big data storage based on DNA", Centre for Real-time Information Networks, Faculty of Engineering and Technology, University of Sydney, Australia, 2015.

[6] Church, G.M., Kosuri, S. and Gao, Y., "Next Generation Digital Information Storage in DNA", Science, 337, 1628, 2012.

[7] Henry, A., Nishanth R. and Panimalar, A., "DNA Digital Data Storage", International Research journal of Engineering and Technology, vol. 5, no 2,2018.

[8] Honwadkar, K. and Laddha, R., "Digital Data Storage on DNA. International Journal of Computer Applications", 142(2), 2016.

[9] Limbachiya, D., Shah, S. and Gupta, M. K., "Dnacloud: A potential tool for storing big data on dna", arXiv preprint arXiv:1310.6992, 2013.

[10]Yang P., Zhong Y.P., Sun D.B., Liu W.Q, Qi J.C., Chen W.Z., Diao W.Y., "Information storage method using synthetic DNA as storage media", China Patent, 2015.