DNA: Future of Digital Storage

Ruchi Kamra[1] Shweta Rana[2]
[1][2]Amity University Haryana

***Abstract*—Human-beings have always been fond of accessing more and more information in minimum possible time and space. Consequently New Generation Computers and High Speed Internet have gained popularity in the recent years. We have been witness to remarkable achievements like the transition from the bulky hard-drives to the flash drives which has made personal data storage efficiently manageable. But when it comes to handling big data, the data of a corporation or of the world as a whole, the present data storage technology comes nowhere near to be able to manage it efficiently. An urgent need for a proper medium for information archival and retrieval purposes arises. Deoxyribonucleic acid (DNA) is seen as a potential medium for such purposes, essentially because it is similar to the sequential code of 0's and 1's in a computer. Seeming to come straight out of science fiction, "a penny-sized device could store the entire information as the whole Internet". The analyzed data from the researches reveals that just four grams of DNA can store all the information that the world produces in a year. Here in this paper we are going to see how DNA can be used as a storage device.**

Keywords: DNA, nucleotides, encryption scheme, storage mechanisms, coding theory, digital data

1. Introduction

In nature, DNA (Deoxyribonucleic Acid) molecules contain genetic blueprints for living cells and organisms is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body has the same DNA. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences. DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. The combination of a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder. An important property of DNA is that it can replicate, or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell.

2. DNA Storage

DNA storage is the process of encoding and decoding binary data onto and from synthesized strands of DNA. To store a binary digital file as DNA, the individual bits (binary digits) are converted from 1 and 0 to the letters A, C, G, and T. The physical storage medium is a synthesized DNA molecule containing these four compounds in a sequence corresponding to the order of the bits in the digital file. To recover the data, the sequence A, C, G, and T representing the DNA molecule is decoded back into the original sequence of bits 1 and 0.

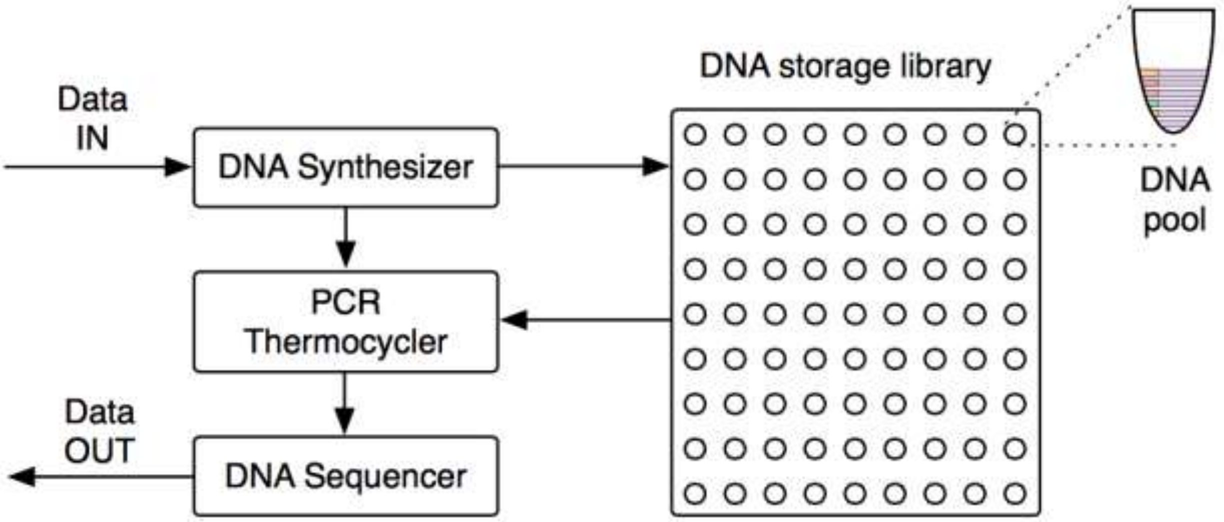

Fig 1 Overview of a DNA storage system

3. Architecture

DNA is used to store data. In this, a delimiter is used at the end of each file so that data can be accessed randomly. The data will be encoded using specialized Huffman tree. If required, each file can be given separate Huffman tree for encoding which will increase data security along with compressing the data. In the case of any error in data while encoding, this error is contained in that file only. As Huffman tree is used for encoding, data compression is achieved. It provides security as no one can decode it without the original tree. For sequencing of DNA strand, a lot of specialized equipments are needed. So without the equipments, DNA cannot be read. Maximum of 2 nucleotide repeats, except for delimiters, are there in DNA. There are 2 copies of all the data. So in the case of data loss, other copy can be used to retrieve data. This method is flexible and the user can manipulate the method to suit the needs and store all kind of data.

3.1 Encoding

1. Form frequency table of characters of the data.

2. Now Huffman tree of non-repeating nucleotides for encoding is generated as follows:

- **a**. Each node in the tree will have 3 children.
- **b**. The weights of branches of children will depend on the incoming weight of parent.
- **c**. If the weight of incoming branch of a parent is A, then C represents the leftmost child, G represents the middle child and T represents the rightmost child.
- **d.** If the weight of incoming branch of a parent is C, then G represents the leftmost child, T represents the middle child and A represents the rightmost child.
- **e**. If the weight of incoming branch of a parent is G, then T represents the leftmost child, A represents the middle child and C represents the rightmost child.
- **f.** If the weight of incoming branch of a parent is T, then A represents the leftmost child, C represents the middle child and G represents the rightmost child.
- **g**. T will be considered to be an incoming weight for r

3. Now split the whole data into overlapping segments of 100 nucleotides with an offset of 50 nucleotides from previous.

4. Form pairs of segments starting from the 1st segment.

5. Index each pair from 0 to 107 and after 107, start from 0 again.

6. Reverse complement 2nd segment in each pair.

7. The index will be of 4 nucleotides long. The index is encoded by a combination of nucleotides in a sequence of A, C, G, T such that no 2 consecutive nucleotides same. Example: 0=ACAC, 1=ACAG, 2=ACAT.

Prepend A and append C to the 1st segment of the pair.

8. Prepend T and append G to the 2nd segment of the pair.

9. Each segment is now synthesized to actual DNA strand of length 106 nucleotides.
If the length of the code of a character is 1, then to avoid repetition of nucleotides, 1 more nucleotide is added in the code of the character.

3.2 Decoding

1. The decoding process is simply the reverse of the encoding process.
2. The 1st nucleotide of DNA will tell whether the . DNA is the 1st or 2nd segment of the pair, whether the data is reverse complemented or not an directionality of strand.
3. If 1st nucleotide is A then:

a. Remove 1st nucleotide
b. Next 4 nucleotides will tell us about segment number.
c. Next 100 nucleotides will be data.
d. The last nucleotide can be used for confirmation of the type of segment.

4. If 1st nucleotide is C then:

a. Reverse whole segment.
b. Remove 1st nucleotide.
c. Next 4 nucleotides will tell us about segment number.
d. Next 100 nucleotides will be data.
e. The last nucleotide can be used for confirmation of the type of segment.

4. Advantages of storing data in DNA

- **They last hundreds of thousands of years.** Yaniv Erlich, a professor at Columbia said that “DNA won’t degrade over time like cassette tapes and CDs, and it won’t become obsolete — if it does, we have bigger problems”. Scientists are able to recover DNA of about 45000 years old. So if we data in data we can retrieve after a long time if needed.
- **455 exabytes of data can be stored in a gram of DNA.** DNA has the capacity to store huge amount of data in small space. Castillo states that all the information in the entire Internet could be located in a device which is lesser than unit cubic inch. So now you can imagine how handy it will be store data for Facebook, Google and many others.
- **Most secure way of storing data** Since DNA is not visible to human naked eye we can’t easily destroy the data. So this becomes most secured way of storage of data.
- **Works with 1/M part of power.** If we start storing and retrieving the data from DNA then the power consumption will be million times less than what our personal computer consumes today to do same work.
- **Increased data writing and retrieving speed.** In DNA data is stored in non linear structure unlike most of the storage devices where data is stored in linear fashion. This allows us to write or retrieve data bidirectional.

5. Challenges

Considering all these major findings, it is inevitable that DNA would become a universal archival medium one day. But it has several challenges, some due to its own physical composition, while some due to technological ineptness to unleash its full potential at present.

The overall process of encoding, amplifying, sequencing, restructuring and decoding takes significantly more time than their conventional counterparts. According to Cox [8], “Assume reading the sequence at enzymatic rates (say 150 nucleotides per second), the retrieval process would still be six orders of magnitude slower than that of a personal computer” (which can read data from the hard drive at nearly 100 Megabits per second). Consequently, DNA is unlikely to compete with optical, magnetic or quantum formats in the foreseeable future.

Many types of errors are associated with the current machines dealing with DNA. For instance, presence of Homo polymers, sequencing errors, error due to lower access rate are some examples. Though DNA in living cells have auto correction enzymes, no such artificial enzymes exist for artificial DNA. DNA Strings need to be discarded if the decoding scheme is inefficient, thus leading to a loss of data and consumption of

more DNA to ensure the same theoretical completeness. Due to its structure, it is prone to mutations in extreme conditions, thus the data might get altered in a mutation. It is a base 4 storage device, so it is fundamentally inefficient since the best storage and lossless compression occurs for base 3 (Huffman Encoding). Another major challenge for practical DNA-based information storage is the difficulty of synthesizing long sequences of DNA de novo (simulating on the computer) to a specified design.
Even with insignificant computational costs and adequate use of the technologies current costs are estimated to be $12,500 per MB for information storage in DNA and $220 per MB for decoding information while that of conventional hard-disk is 8.21 cents (as of 2010).

6. **Conclusion**

Thus, using DNA for data storage, it is possible to store huge amount of data in very less size. As DNA can retain data for millions of years, it is possible to store data for a long time. By using this technique, data is compressed and the security to the data is provided. Parallel reading of files is also possible enabling users to read multiple files at the same time. This technique maintains two copies of data. Hence in case of data damage, its copy can be used to read data. In the case of any errors while encoding the data, the error is restricted to that particular file and no other file is affected due to that error. This technique can be used for all kind of files by making minor changes to adapt to the type of file. This technique can be used to store big data in very small space with little computational overhead. This method is scalable and can be used to store large files too. Also multiple copies can be made easily. This method can be used to store information in archival systems or big data. Instead of using conventional storage devices which have less capacity to store data, DNA- based storage method be used in distant future to store data secured manner and for long time storage and solve the problem of limited space.

References

[1] J. Gantz, D. Reinsel. Extracting value from chaos. International Data Corporation (IDC), Framingham, MA (2011), www.emc.com/collateral/analyst-reports/idc-extracting-value-fromchaos-ar.pdf.

[2] C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland. Long-Term Storage of Information in DNA. Science, 293, 1763 (2001).

[3] George M. Church, Yuan Gao, Sriram Kosuri. Next-Generation Digital Information Storage in DNA. Science, 337, 1628 (2012).

[4] Siddhant Shrivastava and Rohan Badlani. Data Storage in DNA. International Journal of Electrical Energy, Vol. 2, No. 2, June 2014.

[5] Mohan S., Vinodh S. and Jeevan F. R. Preventing Data Loss by Storing Information in Bacterial DNA. International Journal of Computer Applications (0975 8887) Volume 69 No.19, May 2013.

[6] Green, R. E. et al. A draft sequence of the Neandertal genome. Science 328, 710722 (2010).

[7] Willerslev, E. et al. Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science 317, 111114 (2007).

[8] Lilit Garibyan and Nidhi Avashia. Polymerase Chain Reaction. Journal of Investigative Dermatology (2013) 133, e6. doi:10.1038/jid.2013.1

[9] Nick Goldman, Paul Bertone1, Siyuan Chen, Christophe Dessimoz, Emily M. LeProust, Botond Sipos & Ewan Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature, 494, 7780 (2013).

[10] Salunke Avinash N., Shruti Gupta, Varsha Agarwal, and Muhammad Rukunuddin Ghalib. A Novel Digital Information Data Storage Approach in DNA. International Journal of Applied Engineering Research, ISSN 0973-4562, Vol. 8, No. 19 (2013).

[11] Mamta Sharma. Compression Using Huffman Coding. IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.5, May 2010.

[12] Nigam Sangwan. Text Encryption with Huffman Compression. International Journal of Computer Applications (0975 8887), Volume 54 No.6, September 2012.

[13] Ailenberg, M. & Rotstein, O. D. An improved Huffman coding method for archiving text, images, and music characters in DNA. Biotechniques 47, 747754(2009).

[14] Watson, J. D., & Crick, F. H. C. A structure for deoxyribose nucleic acid. Nature 171, 737–738 (1953).

[15] Mohan S, Vinodh S and Jeevan F R. Preventing Data Loss by Storing Information in Bacterial DNA. International Journal of Computer Applications 69(19):53-57, May 2013.