International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

DNA DIGITAL DATA STORAGE

Arockia Panimalar.S[1], Abin Henry[2], Thanga Balu.A[3], Nishanth.R[4]

[1]*Assistant Professor, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Tamilnadu*
[2,3,4] *III BCA A, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Tamilnadu*

Abstract: *Man-kind has always been fond of acquiring more and more information in a minimum possible time. Consequently, new Generation Computers and High Speed Internet have gained popularity in the recent years. We have been witnessing glorious achievements like the change from the bulky hard drives to the flash drives which have made personal data storage easily manageable. But when it comes to handling huge data, the data of a company or of the world as a whole, the present data storage technology comes nowhere near to be able to manage it efficiently. An immediate action needed for a proper medium for information storing and retrieval purposes arises. Deoxyribonucleic Acid (DNA) is seen as a very important medium for such purposes essentially, because it is similar to the sequential code of zeroes and ones in a computer. This field (DNA Computing) has evolved to become a topic of interest for researchers since the past 10 years, with major breakthroughs in its path. Seeming to come straight out of science fiction, "a coin-sized device could store the entire information as the whole Internet". The analyzed data from the researches reveals that just 4 grams of DNA can store all the information that the world can produce in a year. Here this topic of 'Digital Data Storage in DNA' is detailed starting from the first research to the present one, their approach, their advantages and their errors, the need for DNA digital data storage, and how it will ultimately become a paradigm shift in the computing field.*

Key Words: *Nucleotides, Encryption, Decryption, Huffman tree, DNA and Oligonucleotides.*

1. INTRODUCTION

The demand for data storage devices is increasing day by day as more and more data is generated every single day. Total information in digital format in the year 2012 was about 2.7 Zettabytes. Presently devices such as optical discs, portable hard drives, pen drives and flash drives are used to store data. But silicon and the other non-biodegradable materials used in data storage pollute the environment vigorously. Also, they are available in limited quantities. Thus, they would be exhausted one certain day. The linear density of digital storage device is 10 kb per square mm. Hence, newer technology is needed for data storage and storage purposes. As the data increases, the current data storage technology would not be enough to store data in future, as data is growing every single day. Even potentially important information can get lost due to lack of storage. One of the most common causes of data loss is accidental deletion of files without any backup. Every day many people lose their important data because of deleting files accidentally, because they do not have proper backup systems. Poor handling of the optical disk can cause data loss in them. Data loss can happen due to damage to the hard drive. Mechanical damage to the hard drive is common as it contains a lot of damageable parts moving at very high speed. Hard drives can get damaged due to accidental drop of computers. Hard drives can get damaged if any liquid enters it. Liquids can cause damage to electronic parts of drive making it very difficult to recover data. The hard disk can get damaged due to fire accidents. Solid State Drive has a limited number of write cycles. Thus after write cycle limit, it is not possible to write data on them. A printed book has better life expectancy than best of the data storage method.

There are many ways to backup the data. One can use cloud services to store data. But to access data which is stored in a remote cloud, an internet connection is needed all the time. Thus without an internet connection, it is not possible to access the data which is stored in the cloud. Another way is to store data on an external drive. But external drives are prone data loss too. Scientists and Researchers for over the past decade have been trying to develop a robust way of storing data on a medium which is dense, robust and ever-lasting. They are sticking to the storage medium which is used by nature that is DNA. There are many reasons to use DNA as the storage medium such as small size and high density. Just 1 gram of dry DNA can store about 455 Exabyte of data. Thus, data on DNA can be conveniently stored. The power usage required while working with DNA is a very little compared to a conventional storage. Even the error rate of DNA storage is much less than normal storage device.

DNA is a very robust material and it has a long shelf life. The information stored in DNA can be recovered even after thousands of years. As long as the DNA is stored in dry, dark and cold conditions, DNA can be stored for a long time. By using Polymerase Chain Reaction techniques, it is possible to get as many copies as required. Thus, copying of data can be done easily and many copies of data can be obtained. As DNA can retain information for centuries, DNA can be used for long-term storage. Due to high density, the DNA can store a large amount of data in very small space. As in approach, the data is stored in long virtual DNA molecule but encoding is done using synthetically prepared short DNA strand. Short strands will allow to easily manipulating data. It is possible to read simultaneously and randomly read files stored in DNA. Also, compression technique is used to compress data without any loss. The 4 nucleotides of DNA used in the model are Adenine which will be denoted as A, Cytosine as C, Guanine as G and Thymine as T.

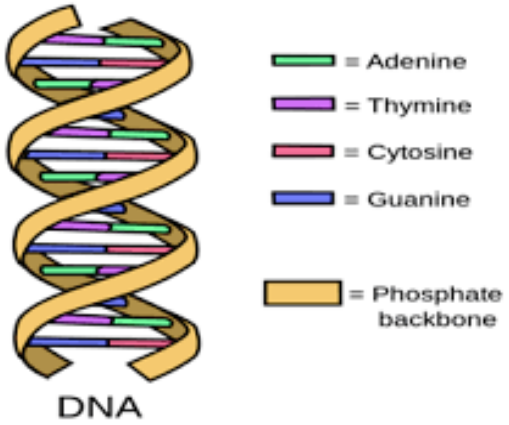

Fig 1: 4 Nucleotides of DNA

2. DNA STORAGE SYSTEM

We imagine DNA digital data storage as the last level of a deep storing hierarchy, giving very dense and durable storage with access times of many hours to days. DNA synthesis and sequencing can be made arbitrarily parallel, making the necessary read and write bandwidths available. We now detail our proposal of a system for DNA based storage with random access support.

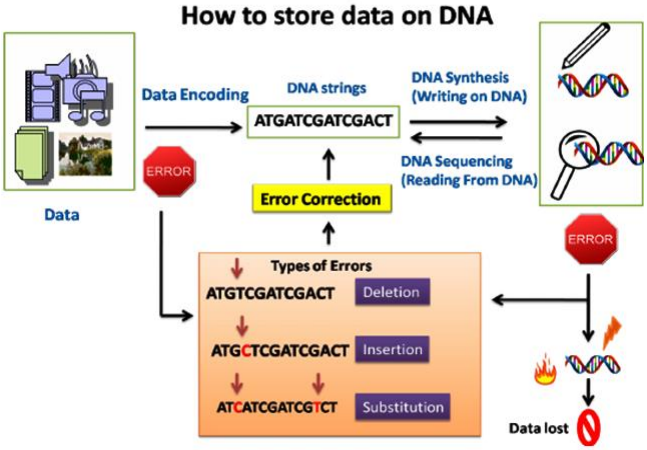

Fig 2: DNA Storage System

3. PROPOSED SYSTEM

In the model proposed in this paper, 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 anyone cannot decode it without the original tree. For sequencing of DNA strand, a lot of specialized equipment is needed. So without the required equipment, 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 kinds of data.

4. WORKING OF DNA DIGITAL DATA

A digital data in DNA should come across 5 levels and are as follows: Coding, Synthesis, Storage, Retrieval and Decoding.

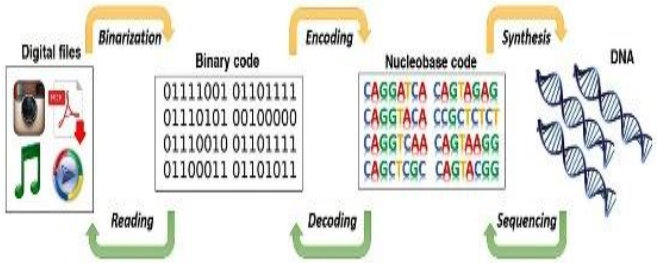

Fig 3: Working of DNA Digital Data

A. ENCODING

1. Form frequency table of characters of the data.

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

- Each node in the tree will have 3 children.
- The weights of branches of children will depend on the incoming weight of parent.
- On the off chance that the weight of incoming branch of a parent is A, at that point C represents to the leftmost child, G represents to the middle child and T represents the rightmost child.
- On the off chance that the weight of incoming branch of a parent is C, at that point G represents the leftmost child, T represents the middle child and A represents the rightmost child.
- On the off chance that the weight of incoming branch of a parent is G, at that point T represents the leftmost child, A represents the middle child and C represents the rightmost child.
- On the off chance that the weight of incoming branch of a parent is T, at that point A represents the leftmost child, C represents the middle child and G represents the rightmost child.
- T will be considered to be an incoming weight for root.

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.

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

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

10. 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.

B. 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 or whether the data is reverse complemented or not and directionality of strand.

3. If 1st nucleotide is A then:

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

4. If 1st nucleotide is C then:

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

5. If 1st nucleotide is G then:

- Reverse whole segment.
- Remove 1st nucleotide.
- Next 4 nucleotides will tell us about segment number.
- Reverse complements next 100 nucleotides.
- These 100 nucleotides will now be data.
- The last nucleotide can be used for confirmation of the type of segment.

6. If 1st nucleotide is T then:

- Remove 1st nucleotide.
- Next 4 nucleotides will tell us about segment number.
- Reverse complements next 100 nucleotides.
- These 100 nucleotides will now be data.
- The last nucleotide can be used for confirmation of the type of segment.

7. If TTTT sequence is found, this will denote the end of the file. The new character will start from next nucleotide.

8. Now by using the same Huffman tree, data can convert the data into original characters. It is possible to generate different Huffman tree for different files or single Huffman tree for whole data. This will compress the data and decoding cannot be done unless one has the original tree. As specific orientation nucleotides have been used in the strands, it is possible to read double number segments in the same number of indexes. The user can read the strand from any direction.

5. EXPERIMENTS

To display the feasibility of DNA storage with random access abilities, we encoded four picture files using the two encodings. The files changed in assess from 5kB to 84kB. We joined these files and sequenced the ensuing DNA to recuperate the files. This section depicts our experience with the synthesis and sequencing procedure, and presents comes about exhibiting that DNA storage is functional and that random access works. We utilized the results of our experiments to illuminate the design of a simulator to perform more experiments exploring the design space of data encoding and durability.

6. MATERIALS AND METHODS

This segment briefly depicts the experimental protocol. We utilized as input to the storage system four picture files. For each picture file(.jpg), we generate DNA sequences.

Fig 4(a):Sydney.jpg, 24301 bytes

Fig 4(b): Cat.jpg, 11901 bytes

Fig 4(c): Smiley.jpg, 5665 bytes

Three image files we synthesized and sequenced for our experiments corresponding to the output of (sydney.jpg, cat.jpg and smiley.jpg). We performed these operations using two different encodings – the Goldman encoding and our proposed XOR encoding. Combined the eight operations produced 45,652 sequences of length 120 nucleotides, representing 151kB of data. To show that DNA storage permits viable arbitrary access, we performed four get activities: choosing three of the four files encoded with the Goldman encoding, and one of the four encoded with the XOR encoding. The three files recovered from the Goldman encoding are in above figure, we likewise performed get (sydney.jpg) on the XOR-encoded data. The synthesized sequences were prepared for sequencing by amplification via the Polymerase Chain Reaction (PCR) method. The product was sequenced using an Illumina MiSeq platform. The selected get operations total 16,994 sequences and 42kB. Sequencing produced 20.8M reads of sequences in the pool. We reviewed the outcomes and watched no peruses of successions that were not chosen – so arbitrary access was powerful in increasing just the objective files.

7. RESULT

We effectively recuperated every one of the four files from the sequenced DNA. Three of the files were recuperated without manual intervention. One file– cat.jpg encoded with the Goldman encoder has acquired a one-byte error in the JPEG header, which we have fixed by hand. As portrayed, the design of the Goldman encoder gives no repetition to the first and last bytes of a file, thus this error was because of arbitrary substitution in either sequencing or synthesis. We could relieve that this error situation by inconsequentially extending that algorithm to wrap the repetitive strands past the end of the file and back to the beginning.

8. CONCLUSION

DNA-based storage has the potential to be the ultimate archival storage solution. It is extremely dense and durable. Thus using DNA for data storage, it is possible to store huge amount of data in very less size. As DNA can hold data for many years, it is conceivable to store data for quite a while. By utilizing this strategy, data is packed and the security to the data is given. 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 can be used in distant future to store data in a secured manner and for a long time storage and can solve the problem of limited space.

9. REFERENCES

[1] J. Gantz, D. Reinsel. “Extracting value from chaos”, International Data Corporation (IDC), Framingham (2011).

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

[3] George M. Church, Yuan Gao, Sriram Kosuri. “Next Generation Digital Information Storage in DNA” (2012).

[4] M. Burrows and D. J. Wheeler, “A block-sorting lossless data compression algorithm,” Digital System Research Center, USA, 1994.

[5]G.C.Smith, C.C.Fiddes, J.P.Hawkins, and J.P.L.Cox, “Some possible codes for encrypting data in DNA”, 2003.

[6] D. A. Huffman, “A method for the construction of minimum-redundancy codes,” 1952.

[7] L. Adleman. Molecular computation of solutions to combinatorial problems. Science, 266(5187):1021–1024, 1994.

[8] 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.

[9] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. Le Proust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," 2013.

[10] C. T. Clelland, et al., "Hiding messages in DNA microdots", 1999.