Next Generation Data Storage (DNA)

Pranjal Bhatt[1] Dhruv Bhatt[2]
[1,2]Department of Computer Engineering
[1]L.D. Engineering College Gujarat, India [2]Dharmsinh Desai Institute of Technology Gujarat, India

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 0s and 1s in a computer. This field (DNA Computing) has emerged to be- come a topic of interest for researchers since the past decade. 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 re- searches reveals that just four grams of DNA can store all the information that the world produces in a year. Here, this topic of Data Storage in DNA is described starting from the very first research to the most recent one, their techniques, their advantages and their flaws, the need for DNA storage.

Key words: Next Generation Data Storage, DNA

I. INTRODUCTION

Data storage and retrieval is inevitable and its preservation problem is looming over our information network. The demand for storing more and more data is increasing day by day. In 2012, the total digital information in the world was about 2.7 zettabytes (10 base 21). With every passing year it is outgrown by its predecessor by 50.

The journey of data storage began from Rocks, Bones, Paper, Punched Cards, Magnetic Tapes, Drums, Films, Gramophone records Floppies etc. Data storage has in the present scenario extended to optical discs including CDs, DVDs, Blu-ray Discs to Portable hard drives and USB flash drives. But all of these techniques are prone to obsolescence and decay. Moreover, Silicon and the other non-biodegradable materials pollute the environment, are limited in resources and would exhaust one day. The Maximum Storage Density on these devices is 1 Terabyte per square inch while the projected data demand would be 8000 Exabytes (10 base 18) till 2015. Libraries, Corporations and File Sharing Systems are in favour of shifting to newer technologies for archival purposes. The current Storage Technologies definitely are not competent enough to handle it efficiently and archive it for the distant future. For instance, the European Organization for Nuclear Research-CERNs CASTOR (CERN Advanced storage manager) system stores 0.08 exabytes of Large Hadron Collider data and grows at 15 PB every year. To store all this information disks are used only for 10 percentage.

DATA, and magnetic tapes need to be used which have access rates reducing every two years. Thus potentially important information is lost for a lack of better archival systems. So to find new solutions to the issues of Digital Data Storage, new technologies and principles are in a state of innovative experimentation throughout the world.

Scientists and Researchers from different parts of the world, over the past decade, have been testing to develop a robust way of storing non-biological data on a medium that is dense, universal, non-obsolete, ever- lasting and enduring. They are sticking to the basics by considering Mother Natures Storage medium, DNA (Deoxyribonucleic Acid). There are several reasons to use DNA as the storage medium. Its storage density and small size (Occupying just 1cm i.e. 1 gram of dry DNA has a storage potential of 455 exabytes of information), something which would take conventional media roughly 2 million times that volume for the same amount of information. Thus data on DNA can be conveniently stored. It can sustain a wider range of temperatures (-800to800C). A gram of DNA contains 1021 DNA bases which can correspond to 108 terabytes of binary data.

The Power Usage while working with DNA is a million times more efficient than a modern personal computer. DNA is a very robust material and has a very long shelf life with no attenuation in data. Data in DNA is stored in a volumetric fashion (using Adenine (A), Thymine(T), Guanine(G), Cytosine(C) bases) which gives access to more storage options unlike present mediums which store data in a linear order. Theoretically DNA can encode 2 bits per nucleotide or 455 exabytes (1018) per gram of DNA. Since the entire sequence never gets damaged during denaturation, the remaining sequence can be amplified to obtain the original one once again. DNA also has the capability for longevity, as long as the DNA is held in cold, dry and dark conditions. It can be suitably protected in a spore, for example, and preserved for millions of years. It can be easily amplified by Polymerase Chain.

One of the most significant advantages of using DNA as a storage medium is that the storage density is very high. For example, it was found in a research by Hoch and Losick that the density to contain characters (char/m2) of a Bacillus subtilis bacterium (genome size 4.2 Mega Base Pairs, with 1 m diameter) spore is twenty million times that of a 200 Megabyte ZIP disk of diameter 10 cm. DNA sequences can contain more information than their binary counterpart because DNA with 4 bases has 4X representations possible for a X- character long string while binary system can represent only 2-X times that information.

II. Storage mechanism and encoding schemes

The DNA consists of double stranded polymers of four different nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). The primary role of DNA is long-term storage of genetic information. This feature of DNA is analogous to a digital data sequence where two binary bits 0 and 1 are used to store the digital data. This analogous nature of DNA nucleotide with Binary Bits can be exploited to use artificial nucleotide data memory. For example, small text message can be encoded into synthetic nucleotide sequence and can be inserted into genome of living organisms for long term data storage. Further, to enhance the data density for encoded message, original text message can be com- pressed prior to encoding. Currently, there exist many losses-less compression algorithms for large text les. All of them need sufficient context information for compression, but context information in small le (50 kB to100 kB) is difficult to obtain. In small les, context in- formation is sufficient context information only when we process them by characters. Character based compression is most suitable for small les up to 100 kB. Thus we need a good compression algorithm, which re- quires only small context or we need an algorithm that transforms data into another form. An alternative approach is to use Burrow Wheeler transform followed by Move to Front transform. The human encoding is used to convert the original file into compressed one. There has been much advancement in the use of DNA as a data storage device. One of the most critical steps in the realization of biological data storage is the conversion of digital data to nucleotide sequence.

Scientist proposed the idea of using hereditary media as a media for information transmission in communication process. Shuhong Jiao devised a code for DNA based cryptography and stegano-cryptography and implemented in artificial component of DNA. Nozomu Yachie used keyboard scan codes for converting the in- formation to be encoded into hexadecimal value and finally binary values. The last step was to translate the bit data sequence into four multiple oligo nucleotide sequence. This was mapped with the nucleotide base pairs. Chinese University of Hong Kong used Quaternary number system to transform the information for mapping it to nucleotides. First they obtained ASCII vale of the information and used the mapping table 0=A, 1=T, 2=C and 3=G for the formation of nucleotide strand. In this method of encoding nucleotides the number of binary bits used for representing the digital in- formation was same as the nucleotide strand. Storing information in cells overcomes the constraints of disk such as: rotational delay, disk speed, and a real density because there are no moving mechanical parts in a cell. Cells hold DNA information. DNA is the information that cells use to create life and make proteins. A DNA strand might look like the following:

TAT GCCT GACGGAA
ATACGGACTGCCTT

(A − Adenine, T − Thymine, G − Guanine, C − Cystine)

These amino acids always bind or hybridize in the following order:

$A -> T$
$C -> G$
$T -> A$
$G -> C$

The Crick-Watson double stranded model provides a good example of the four nucleotides that make up the DNA model (A, T, C, and G) and how the four base pairs hybridize, or join together.

The creation of a protein starts with transcription, the copying of DNA into mRNA. Proteins are synthesized using mRNA as a template. In an abbreviated form, mRNA goes through translation, the final process in which a protein is created. In translation, the binding changes slightly:

$A -> U$
$C -> G$
$U -> A$
$G -> C$

Thymine (T) is replace with Uracil (U) another base. A piece of mRNA after going through translation might look like the following (single stranded model):

UCU GUUUUACGGCCCAUUAA

To create a protein, a codon or nucleotide, forms to create an amino acid. Looking at Table 1, the Genetic Code Table, the first codon (UCU) would create the amino acid Ser. The second codon (GUU) would create Val. In order to create words, assign letters to amino acids. As amino acids are formed, words are formed. For demonstration purposes, the following letter have been assigned:

Ser = N
Val = A
Ser = N
Glu = O

III. Data interpretation

A. Context Information Generation

Currently there are many compression methods that require good context in order to achieve a good compression ratio. One of them is Burrow Wheeler transform. BWT can achieve good compression ratio provided that there is a sufficient context which is formed by frequent occurrence of symbols with same or similar prefixes. Maximizing the context leads to better compression ratio. The Burrow Wheeler algorithm is based on the frequent occurrence of symbol pairs in similar context and it uses this feature for obtaining long strings of the same characters. These strings can be transformed to another form with move to front (MTF) transformation.

B. Compression of Text File

We used statistical compression method to com- press the data obtained after transformation. The chosen statistical compression scheme was human encoding. Input consists of alphabet A and set W represented in equation (1) and (2) respectively. Output is a set of binary sequence in equation (3), which must satisfy the goal (4) for all the codes with the given condition.

A = a1, a2, .., an
W = w1, w2, .., wn
wi = weight (ai); 1 < i < n
C(A,W) = c1, c2, .., cn

IV. MAPPING FUNCTION

A. Mapping Table

Mapping table consists of binary bits and nucleotides. Binary value is represented as 0 and 1.Nucleotides are represented as A, C, G and T. Four binary bits are represented by two nucleotide base pairs resulting in sixteen such combinations as shown in Mapping Table. The reason for choosing four bits for two nucleotides is that the output of human encoding International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.2, April 2012 give the mapping of Hexadecimal value (radix =16). So we need sixteen such combinations to represent this in binary and then nucleotides.

Binary nts:	*Binary nts:*	*Binary nts:*	*Binary nts:*
0000-AA	0100-AC	1000 -AG	1100 - AT
0001-CA	0101 -CC	1001 - CG	1101 - CT
0010 -GA	0110 -GC	11010 - GG	1110 - GT
0011 -TA	0111 - TC	1011 - TG	1111 - TT

Table 1: Mapping table

V. RECALLING DATA

A mass spectrophotometer (MS) is used to identify compounds, or in our case proteins, based on their mass and charge [2]. Data can be recalled and displayed by using a mass spectrophotometer. As the protein, or polypeptide chain is created, it can be analysed using a MS to read the amino acids one peptide at a time [3]. Figure 1 is an example of the output from a mass spectrophotometer. Each spike is an amino acid. You can easily identify the amino acid by its value. If each amino acid is assigned a character, you would read the words by reading the amino acids.

A. Erasing Data

All cells are subject to environmental and growth conditions. These include, but are not limited to: temperature, pH, growth nutrients, oxygen, and CO2. If any one of these gets out of balance, it could kill a cell. To erase data, the pH could be altered and the cell would die and the data along with it. Unlike traditional disk based storage, once the cell is dead it cannot be reread.

B. Changing Data

One tool that biotechnologists use to manipulate a protein is site directed mutagenesis. This tool would allow a technician to change specific base pairs causing a change to amino acids, in our case data, to whatever they need to [4]. This has the same affect as an editor using a word processor. Site directed mutagenesis also allows us to insert and delete base pairs as needed. This technique requires a DNA primer containing the desired change. This primer has to hybridize with the DNA that contains the gene of interest. The fragment is extended and copied. The fragment is then introduced into a cell to be cloned.

C. Compression

Proteins are not nice long straight chains. After they form, they fold up into a shape. This is called con- formation. All proteins of the same type fold the same way [5]. This folding naturally acts like compression. The protein cannot be read if it is folded. To unfold it, the protein must be heated. When allowed to cool, the protein will return to its conformation [6].

D. Copying Data

Systems Administrators create redundant copies of data to protect against possible data loss. Cells reproduce naturally by cell division. When a cell divides, it copies its DNA into the new cell. If a second copy of data is needed before the cell divides, DNA can be replicated by using Polymerase Chain Reaction (PCR), a process used to amplify DNA.

E. Error Rates

The error rate during cell division is approximately 1 in 100,000 nucleotides. For a human, this would be unacceptable. The cell can correct these pre-replication errors by a process called proofreading. This proofreading process can fix up to 99 errors, up to 99.99 mismatch repair, a post-replication process. Errors occur naturally during replication. A.01 percent error rate should be more than acceptable for most processing requirements.

VI. MESSAGE ENCODING AND RETRIEVAL

We have implemented the data encoding in nucleotides by integrating the trans-formation algorithm with statistical compression scheme. Here we have demonstrated our encoding and message retrieval scheme on small text: OPERATION BAR- BAROSSA. The first step was to perform Burrow wheeler transform and move to front transform on the original text. This was done to generate better context information and obtain high compression ratio. The security of the encoded message was maintained by encryption method. The encryption method used was One Time Pad where a randomly generated binary strand was XORed with the binary strand obtained from human encoding. We used a random function generator for generating the random binary sequence.

Only in the last step, Huffman encoding method was introduced which compressed the original text message to a much smaller size. The next step toward message encoding was to use mapping table. The generated binary strand obtained after Huffman encoding was mapped to nucleotides according to the mapping table.

Second phase of our work was to convert the encrypted binary strand into nucleotide sequence. Al- though many other mapping functions can be used, but for our convenience we used two nucleotides to represent four binary bits, as hexadecimal (radix =16) value is being converted to four bit binary representation and thus leading to formulation of original text message in form of nucleotide sequence.

The decoding of message can be performed by re- versing the encoding scheme. This is explained in Figure 1. This nucleotide sequence can be artificially synthesized and inserted into the host to maintain the attributes of hereditary media and durable data storage for intensive period of time. We have not proceeded in implementing the biological protocols to insert the sequence in genome of bacteria.

We describes a data encoding method to achieve high volume data density by reducing the number of nucleotides. The primary focus of this study was to encode data for les (light emitting surface) of very small size.

Data encoding method was performed into two steps. 1. This step was to compress the original text message. This was achieved using transformation and compression algorithm. 2. Introduction of mapping table, which finally maps the binary strand to nucleotide sequence.

VII. CHALLENGES

Considering all these major findings, it is inevitable that DNA would become a universal archival medium one day. But it presents several challenges, some due to its own physical composition, while some due to our 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 scientist 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 Strings need to be dis- carded 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 de- sign.

Even with insignificant computational costs and adequate use of the technologies current costs are estimated to be 12,500 dollar per MB for information storage in DNA and 220 dollar per MB for decoding in- formation while that of conventional hard-disk is 8.21 cents (as of 2010).

VIII. CONCLUSION

It is clear that data storage in DNA is no more con- fined to science fiction but is being realized and improvised at very promising rates by research teams all over the world. This idea has received positive criticism from the general public as can be inferred from their responses on the different science websites. Similar to all revolutions in technology, DNA-based data storage technology has to face major challenges to realize its full potential. It is however, inevitable that DNA would be invariably used for archival purposes for its sheer density, robustness, stability and energy efficiency. In theory, grams of DNA can store all the information ever produced by mankind. Several breakthroughs will be required before it becomes commercially mainstream for data retrieval.

This field has had a million-fold improvement in the recent years. Digital Data Storage in DNA technology shows immense progress, since reading and writing it is advancing ten times every year unlike the Electronic Technology which is improving roughly 1.5 times a year (Moore’s Law).

Cost efficiency also shows promise as the exponential drop in DNA synthesis and sequencing cost has been five-fold and twelve-fold respectively while Electronic media show only a 1.6-fold drop per year. As handheld sequencers are becoming commercially avail- able, the research would grow beyond large projects, with more people getting access to experiment with this idea. As DNA is the basis of life on Earth, methods for working with it, storing it and retrieving it will remain the subject of continual technological innovation, says Nick Goldman. In order to handle big data a practical large-scale DNA archive would need stable DNA management and innovative indexing solutions. This would create the desired paradigm shift in computing as the aspect of Data Storage would be an integral part to realize the idea of DNA Computing. Reciprocally, it would catalyze research and development in synthesis and sequencing technologies.

In 2015 Microsoft successfully store 1MB of data in small jar of DNA, and now their target is to store 1GB of data storage in DNA so it is almost 1000 times of previous storage.

IX. Future enhancement

Future work could include compression schemes; dealing with redundancy at all levels, checking for parity, correcting errors to enhance density and safety. DNA could also be substituted with polymers or be modified to suit the needs of digital storage. Furthermore, it will fuel research to look for alternative materials for information storage and to aid in realizing the need for a universal medium for data. Overall, this technology is here to stay and could transform the way we have ever looked at DNA and computing as totally different entities.

References

[1] Thompson, D.A. Best, J.S. (n.d.). The future of magnetic data storage technology. IBM Journal of Research and Development. Vol. 44, No. 3, Pg. 311

[2] (2009) Basic laboratory methods for biotechnology. Pearson Education, pp 435

[3] Mullen, D.L. (n.d.), Ionization Methods in Mass Spectrometry. School of Chemical Sciences, University of Illinois.

[4] Zheng, L, Baumann, U., Reymond, J.L..(2004). An efficient one step site-directed and site- saturation mutagenesis protocol. Nucl. Acids Res. 32:e115, 2004

[5] Ross, G., Fleming, P., Banavar, J., Maritan, A (2006). A backbone based theory of protein folding. Proc. Natl. Acad. Sci. U.S.A. 103 (45): 16623-33.

[6] H.Goto, Y. Hasegawa, Tanaka, M. (2007). Effi- cient Scheduling Focusing on the Duality of MPL Representatives. Proc. Symp. Comput. Intel. In Scheduling.