Digital Data Storage on DNA

A Natural Computing Project report submitted

by

Aman Chopra 140911358

Csaradhi Tejas 140911310

Tanveer Sapra 140911266

Vibhavari 140911013

Department of

Information and Comunication Technology,

MIT, Manipal

MANIPAL
INSTITUTE OF TECHNOLOGY
A Constituent Institute of Manipal University, Manipal

October 2017

ABSTRACT

Digital data has changed the use and access of information. Everyday lot of data is produced and this requires high- density storage devices which can retain values for a long time. One of the major problems faced today in data storage is the data loss due to crashing of storage devices like magnetic disc or optical disc. In recent years scientists have turned their attention towards the bio materials for data storage. Attempts were made to store data in proteins, tissues, etc. But the problem faced over these things are, that they are not reliable, easily biodegradable and information cannot be transmitted over generations [1].

DNA can be potentially used for these purposes as it is not much different from the conventional method used in a computer. DNA can be used as a robust and high-density storage device even under unfavourable conditions. Data stored in DNA is more reliable than any other devices. This invention provides a hope for future data storage which will be safe for millions of years, as they are transmitted over various generations. We can cultivate the bacteria in which the data are stored in various places like fish tank, garden, stomach, etc.

DNA has many potential advantages as a medium for immutable, high latency information storage needs. For example, DNA storage is very dense, unlike most digital storage media, DNA storage is not restricted to a planar layer, and is often readable despite degradation in non-ideal conditions over millennial, finally, DNAs essential biological role provides access to natural reading and writing enzymes and ensures that DNA will remain a readable standard for the future.

Keywords: DNA, data storage, encoding, compression, storage mechanism, digital data, secured storage, Huffman tree, Bacterial storage.

Contents

List of Tables

List of Figures

ABBREVIATIONS

ASCII	:	American Standard Code for Information Interchange
CD	:	Compact Disc
DNA	:	Deoxyribonucleic acid
EMBL	:	European Molecular Biology Laboratory
GB	:	Gigabytes
JPG	:	Joint Photographic Experts Group
mm	:	millimetre
nm	:	nanometre
OLS	:	Oligo Library Synthesis
ssDNA	:	Single Stranded Deoxyribonucleic acid
UV	:	Ultraviolet

NOTATIONS

A	:	Adenine
T	:	Thymine
G	:	Guanine
C	:	Cytosine
nt	:	Nucleotide
°C	:	Degree Celsius

Chapter 1

Introduction

1.1 DNA as a Storage Medium

One of the major researches in data storage is done in Bio Storage Technologies. It refers to encoding digital information on a biological molecule. It is half Information Technology oriented and the rest Biotechnology. Earlier, we have discovered storing data on protein, blood tissue, etc. It is not reliable as the biological material can degrade over time. In order to overcome this defect, we store data in DNA, by which data can be transmitted from one generation to another. So we go for storing data on genetic material which can be transmitted for several generations.

The demand for data storage devices is increasing day by day as more and more data is generated every day. Total information in digital format in the year 2012 was about 2.7 zettabytes. Presently devices such as optical discs, portable hard drives, and flash drives are used to store data. But silicon and the other non-biodegradable materials used in data storage pollute the environment. Also, they are available in limited quantities. Thus, they would be exhausted one day. The linear density of digital storage device is 10 kilo bites per square mm. Hence, newer technology is needed for data storage and

archival process. As the data increases, the current data storage technology would not be enough to store data in future as data is growing every day. Even potentially important information can get lost due lack of storage space.

DNA storage is hierarchical and multi-layer. We may say it consists of files grouped into folders which are organized into volumes of data. Every animal cell has a nucleus (the central and controlling part of the cell). This contains some fixed amount of chromosomes (depending upon the organism). These chromosomes have many genes. Each gene is made up of millions of DNA. These nucleic acids are just complex organic molecules. As per Watson and Crick (persons who designed the DNA structure at first) it is a helical ribbon structure. There are two ribbons on either side, each contains alternatively placed sugar and protein molecules. These proteins are bonded with Adenine, Guanine, Thiamine and Cytosine.) Any small change in the environment can easily affect this molecule which may sometimes undergo mutation. Mutation is a small or may be even bigger sudden change in the genome sequence. This may even change the entire organism. This mutation helps us in data storage [2].

The DNA is as stable as its living carrier (bacteria, plant or animal cell). In some cases after the cells death the DNA data may survive. In suitable conditions the DNA may hold its data for years. On the other hand the DNA is vulnerable to hydrolysis and oxidation. Many factors lead to mutations in the genes.

DNA is a very good content that contains lots of genetic material and good natural biological data storage medium. It contains all the information about an organism in its complex structure figure 1.1. The organisms complete details like the physical structure, ancestral information, capability, etc.

are naturally stored in this type of materials. Any change in it can change the complete nature of the organism. But some part of DNA is not used throughout its life. Such portion of DNA is identified and data is stored in it.

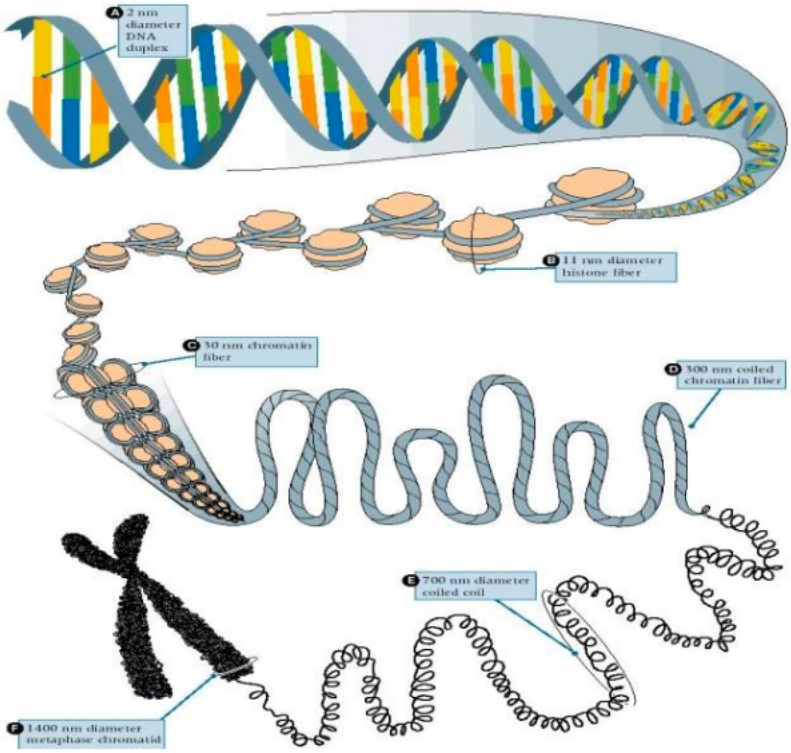

Figure 1.1: DNA Structure

Chapter 2

Literature Survey and techniques

Researchers from the Japanese universities, Keio University Institute for Advanced Biosciences and Keio University Shonan Fujisawa Campus in Tokyo have succeeded in creating an artificial DNA containing the data to be preserved. The study, published in the academic journal Biotechnology Progress, describes how DNA encoded with "E=MC2 1905!" was put into the genome of the common soil bacterium Bacillus subtilis. While ink may fade and computers may crash, bacterial information lasts as long as a species stays alive possibly a mind-boggling million years.

In this report, we have considered four different research works done in the field of Data Storage in DNA. We will discuss the algorithms used, the advantages and applications. The works are arranged in the chronological order.

2.1 Next-Generation Digital Information Storage in DNA

This paper is by George M. Church, Yuan Gao and Sriram Kosuri from Harvard Medical School Boston. It was published on August 2012.

2.1.1 Introduction

Storing messages in DNA was first demonstrated in 1988 and the largest project to date encoded 7920 bits. The small scale of previous work stems from the difficulty of writing and reading long perfect DNA sequences, and has limited broader applications and hence in this paper, a novel approach is proposed to encode digital information in DNA.

2.1.2 Proposed Method

In this method, authors have developed a strategy to encode arbitrary digital information using a novel encoding scheme that utilizes next-generation DNA synthesis and sequencing technologies. They converted an html-coded draft of a book that included 53,426 words, 11 JPG images and 1 JavaScript program into a 5.27 megabit bitstream. They then encoded these bits onto 54,898 159nt oligonucleotides (oligos) each encoding a 96-bit data block (96nt), a 19-bit address specifying the location of the data block in the bit stream (19nt), and flanking 22nt common sequences for amplification and sequencing. To read the encoded book, they amplified the library by limited-cycle PCR [3]. They joined overlapping paired-end 100nt reads to reduce the effect of sequencing error. Then using only reads that gave the expected 115-nt length and perfect barcode sequences, they generated consensus at each base of each data block at an average of 3000-fold coverage. All data blocks were recovered with a total of 10 bit errors out of 5.27 million which were predominantly located

within homo-polymer runs at the end of the oligo where they only had single sequence coverage.

2.1.3 Advantages

1. They encoded one bit per base (A or C for zero, G or T for one), instead of two. This allows them to encode messages many ways in order to avoid sequences that are difficult to read or write such as extreme GC content, repeats, or secondary structure. By splitting the bit stream into addressed data blocks, they eliminate the need for long DNA constructs that are difficult to assemble at this scale.

2. To avoid cloning and sequence verifying constructs, they synthesize, store, and sequence many copies of each individual oligo. Since errors in synthesis and sequencing are rarely coincident, **each molecular copy corrects errors in the other copies**. They use a purely in vitro approach that avoids cloning and stability issues of in vivo approaches.

3. they leverage next-generation technologies in both DNA synthesis and sequencing to allow for encoding and decoding of large amounts of information for 100,000-fold less cost than first generation encodings.

2.1.4 Results

The proposed method was found to be better than other approaches prevalent at that time. They plotted information density ($\log_{10}$ of bits/mm3) versus scalability as unit and the result is shown in graph figure 2.1. This general approach of using addressed data blocks combined with library synthesis and consensus sequencing should be compatible with future DNA sequencing and synthesis technologies. Reciprocally, large-scale use of DNA such as for information storage could accelerate development of synthesis and sequencing technologies.

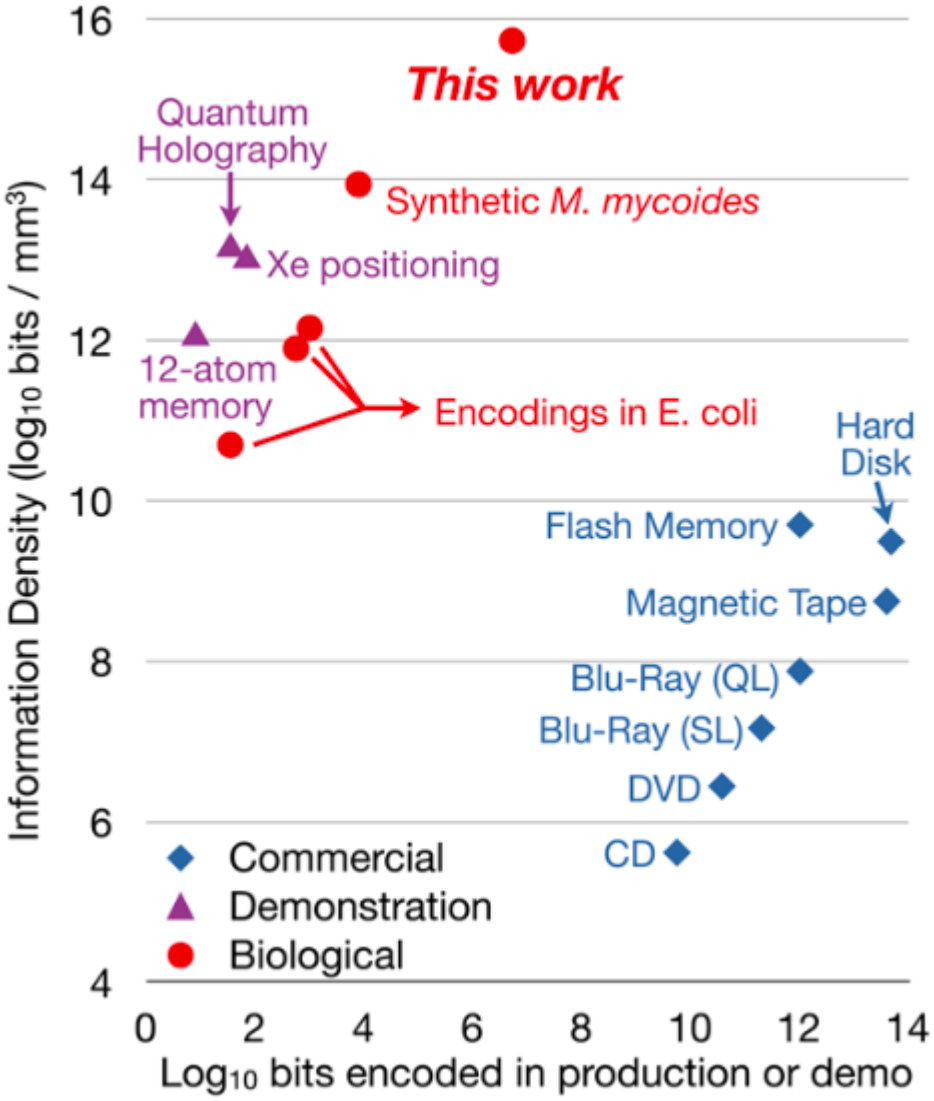

Figure 2.1: Comparison to other measured by the $\log_{10}$ of bits encoded in the report or commercial technologies

2.2 Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

This paper is by Nick Goldman and Paul Bertone.

2.2.1 Introduction

The challenge of data outburst has focused some interest on DNA as an attractive target for information storage1 because of its capacity for high-density information encoding, longevity under easily achieved conditions and proven track record as an information bearer. Previous DNA-based information storage approaches have encoded only trivial amounts of information or were not amenable to scaling-up, and used no robust error-correction and lacked examination of their cost-efficiency for large-scale information archival. Here we describe a scalable method that can reliably store more information than has been handled before. We encoded computer files totalling 739 kilobytes of hard-disk storage and with an estimated Shannon information10 of $5.26 * 10^6$ bits into a DNA code, synthesized this DNA, sequenced it and reconstructed the original files with 100% accuracy. Theoretical analysis indicates that our DNA-based storage scheme could be scaled far beyond current global information volumes and offers a realistic technology for large-scale, long-term and infrequently accessed digital archiving.

2.2.2 Proposed Method

Although techniques for manipulating, storing and copying large amounts of existing DNA have been established for many years, one of the main challenges for practical DNA-based information storage is the difficulty of synthesizing long sequences of DNA to an exactly specified design.

As a proof of concept for practical DNA-based storage, we selected and en-

coded a range of common computer file formats to emphasize the ability to store arbitrary digital information. The following steps are followed:

1. The bytes comprising each file were represented as single DNA sequences with no homopolymers.

2. Each DNA sequence was split into overlapping segments, generating four-fold redundancy, and alternate segments were converted to their reverse complement as shown in figure 2.2.These measures reduce the probability of systematic failure for any particular string, which could lead to uncorrectable errors and data loss.

3. Each segment was then augmented with indexing information that permitted determination of the file from which it originated and its location within that file, and simple parity-check error-detection.

4. They synthesized oligonucleotides (oligos) corresponding to the designed DNA strings using an updated version of Agilent Technologies OLS (oligo library synthesis) process [4].

5. After resuspension, amplification and purification, we sequenced a sample of the resulting library products at the EMBL Genomics Core Facility.

6. Strings with uncertainties due to synthesis or sequencing errors were discarded and the remainder decoded using the reverse of the encoding procedure, with the error-detection bases and properties of the coding scheme allowing us to discard further strings containing errors.

7. Full-length DNA sequences representing the original encoded files were then reconstructed in silico. The decoding process used no additional information derived from knowledge of the experimental design.

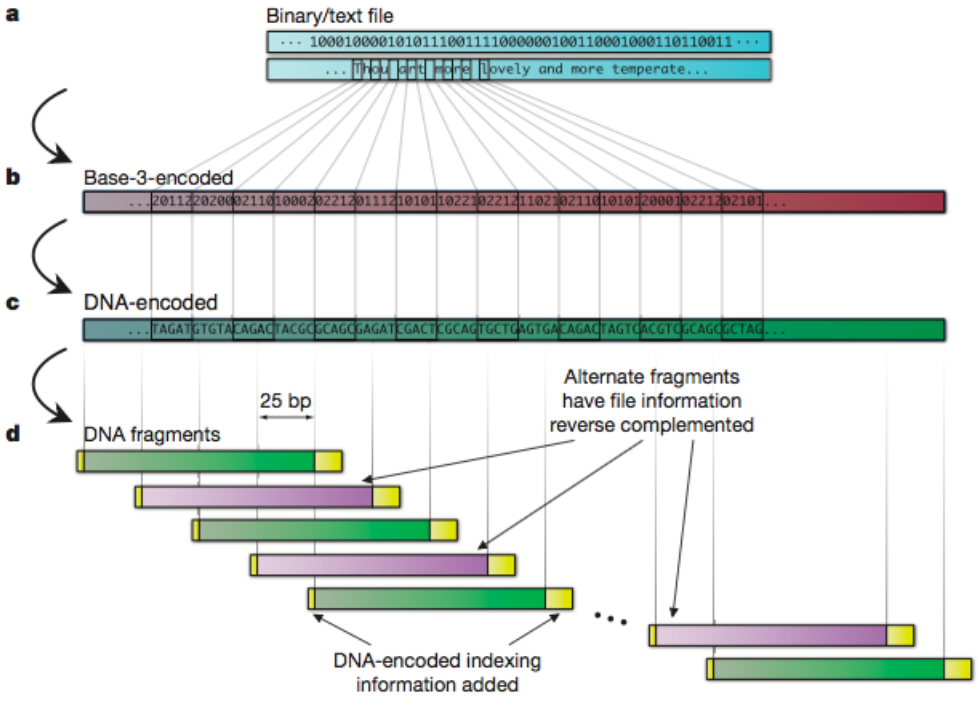

Figure 2.2: Digital information encoding in DNA

An important issue for long-term digital archiving is how DNA- based storage scales to larger applications. The number of bases of synthesized DNA needed to encode information grows linearly with the amount of information to be stored, but we must also consider the indexing information required to reconstruct full-length files from short fragments. As indexing information grows only as the logarithm of the number of fragments to be indexed, the total amount of synthesized DNA required grows sub-linearly and as the total amount of information increases, the encoding efficiency decreases only slowly as shown in figure 2.3.

2.2.3 Advantages

Extension of scaling analysis to model the influence of reduced sequencing coverage on the per-decoded-base error rate revealed that error rates increase only very slowly as the amount of information encoded increases to a global data scale. This also suggests that the mean sequencing coverage of 1,308

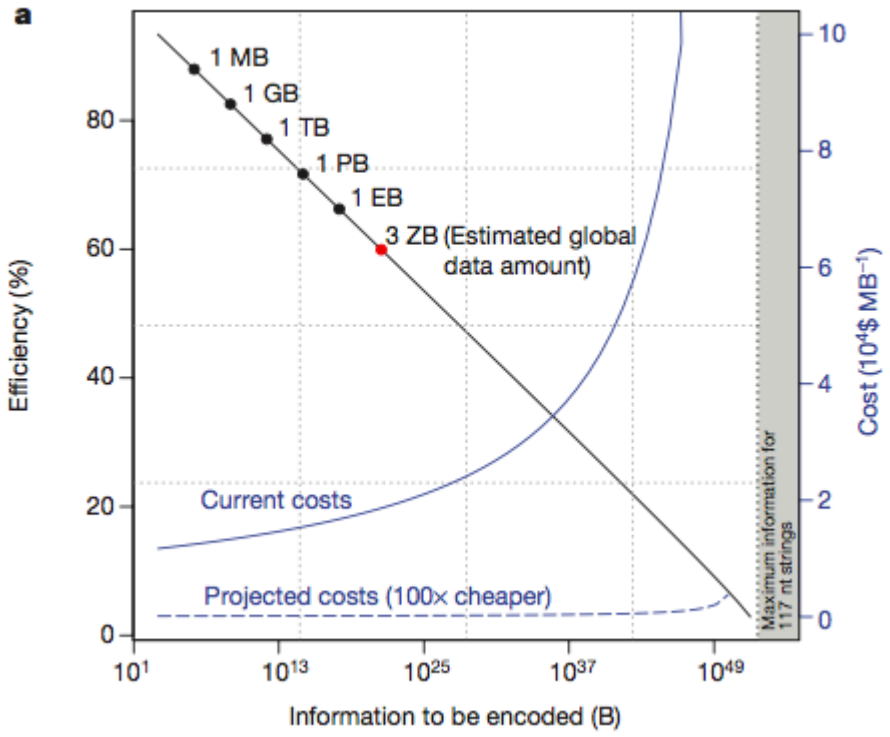

Figure 2.3: Scaling properties and robustness of DNA-based storage

times was considerably in excess of that needed for reliable decoding. Authors confirmed this by sub sampling from the 79.6 3 106 read-pairs to simulate experiments with lower coverage. Reducing the coverage by a factor of 10 (or even more) would have led to unaltered decoding characteristics, which further illustrates the robustness of this DNA-storage method. DNA-based storage might already be economically viable for long- horizon archives with a low expectation of extensive access, such as government and historical records.

2.2.4 Results

This sequencing protocol consumed just 10% of the library produced from the synthesized DNA, already leaving enough for multiple equivalent copies. Existing technologies for copying DNA are highly efficient meaning that DNA is an excellent medium for the creation of copies of any archive for transportation, sharing or security. They find that with current technology and

this encoding scheme, DNA-based storage may be cost-effective for archives of several megabytes with a ,600 5,000 year horizon [5]. One order of magnitude reduction in synthesis costs reduces this to 50500 year; with two orders of magnitude reduction, as can be expected in less than a decade if current trends continue.

2.3 Preventing Data Loss by Storing Information in Bacterial DNA

This is by Mohan S and Vinod S from PSG College of Technology, Coimbatore.

2.3.1 Introduction

DNA (Deoxyribo Nucleic Acid) is a very good content that contains lots of genetic material and good natural biological data storage medium. It contains all the information about an organism. The organisms complete details like the physical structure, ancestral information, capability, etc. are naturally stored in this type of materials. Any change in it can change the complete nature of the organism. But some part of DNA is not used throughout its life. Such portion of DNA is identified and data is stored in it. Now we are choosing bacteria DNA for data storage.

The DNA is as stable as its living carrier (bacteria, plant or animal cell). In some cases after the cells death the DNA data may survive. In suitable conditions the DNA may hold its data for years. On the other hand the DNA is vulnerable to hydrolysis and oxidation. Many factors lead to mutations in the genes. In mutations, letters of the genetic code can be changed and stretches of DNA can be deleted.

2.3.2 Proposed Method

2.3.2.1 Mutation by UV Rays

When the ultraviolet rays fall on the DNA it absorbs some energy. This energy excites the electrons in the DNA to a higher energy level. As a result it results in mutation as shown in figure 2.4. The two nearby thiamine molecules get bonded with each other leaving the adenine molecule. This forms an intercalating agent in the DNA, which is unstable. This instability causes the DNA to get muted. Again when it gets into the normal state the genome sequence is changed and results in mutation [6]. This is induced by ultraviolet rays. The role of ultraviolet rays is to increase the mutation rate. Now here the authors control the mutation rate by digitally made electrical signals. This forms the basic principle in data storage in DNA.

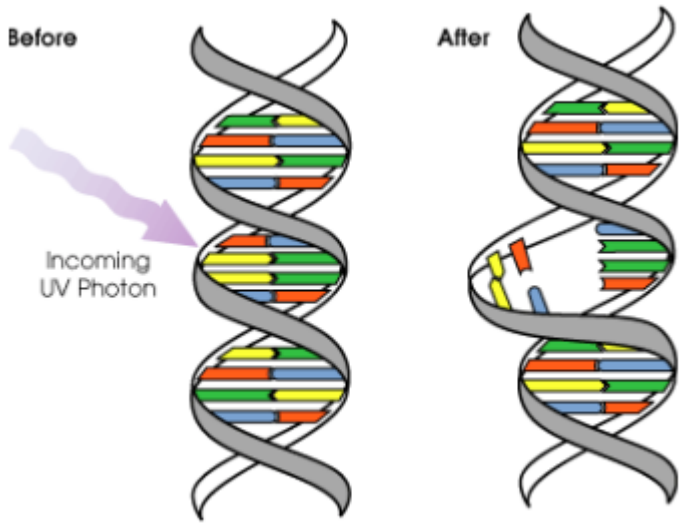

Figure 2.4: DNA in Data Acceptance form

2.3.2.2 Encoding

1. At first a suitable bacterium is chosen which must be chosen in a way that it must tolerate high temperatures, desiccation, ultraviolet light and ionizing radiation doses 1000 times higher than would be fatal to

humans. Bacteria like Escherchia coli, Deinococcusradiodurans, Serratiamarcescens, Bacillus subtilis, etc are normally used as they can tolerate extreme conditions.

2. The DNA molecule is extracted from it by Chelexs process (A well known biotechnological method to collect bacterial DNA on water on a test tube). Then it is stored in a proper temperature.

3. Then the DNA is taken for Bioluminescence. It is a process of taking the DNA to a state by which it is ready to accept electrical signals.

4. In the presence of ultraviolet light the liquid medium containing DNA is encoded with digital data. The presence of ultraviolet light is to increase the effect of mutations. By the effect of ultraviolet light the effect of mutation becomes more spontaneous.

5. Now two electrodes are inserted into this liquid medium which contains the DNA as shown in figure 2.5.

6. The digital data is converted into electrical impulses. These impulses are sent through these electrodes and this will encode the data onto the DNA. After encoding the data on the synthetic DNA, the DNA is inserted into living bacteria as a living medium. This data will live along with the bacteria for generations.

2.3.2.3 Decoding

Data stored on the DNA can be retrieved by decoding it. The reverse process in the encoding is just done. The DNA is extracted from the bacteria at first. The luminometer is again added and the concentration of DNA is fairly reduced. Electrodes are placed into the liquid medium that contains DNA molecules and ultraviolet light and ionizing radiations are given to the content

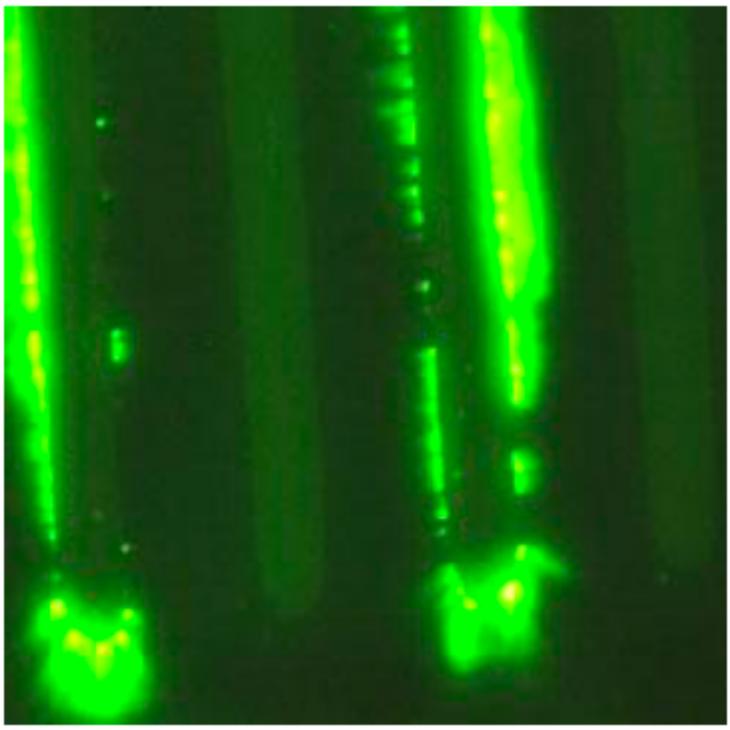

Figure 2.5: The gold rods that is inserted in the DNA content

at 30°C. The encoded data in the DNA get transformed into the electrical impulses and the data are retrieved through the electrodes. There will not be any loss of information due to this process. The original information that was encoded is retrieved from the content. The amount of information that can be stored depends up on the amount of bacteria in the content. If a milliliter of liquid can contain up to billion bacteria, the potential capacity of such a memory system is enormous. Higher the bacteria content, higher the storage capacity.

2.3.3 Advantages

By this method data can be stored in the DNA and data will be transmitted over many generations. These data can last even up to 3000 years. Hence, even after 100 human generations we can get the data. Bacteria may be an inexpensive and stable long-term means of data storage. We can store high

content of data on a small piece which can just measure a few nanometers. Genetic coding is so massive that information can be stashed away somewhere in the gene without affecting an organism's overall appearance and other traits. It can be used in many applications like:

1. This biomaterial can be made in small memory sticks that can store data. By this the todays optical memory storage can be replaced by Biostorage Technologies.

2. In the optical memory system the range is band limited by the wavelength of light. But here we dont have any band limitations. So we can store enormous amount of data without any band limitations.

3. We can spot the sense of living organisms in places. This will be helpful to identify bodies struck in places due disasters.

4. Other applications are to create living data storage for nano-computers, or to tag an organism for life using a unique identifier strand inserted into the DNA. Computer memory like the magnetic disc, semiconductors, superconductors, etc may fail due to lots of problems. It can be affected by external radiations. But the data stored here is more reliable.

2.3.4 Results

This simple, flexible and robust method offers a practical solution to data storage and retrieval challenges in combination with other previously published techniques. One early use for the technique would be to create special markers to identify legitimate versions of pharmaceuticals. However, the bacillus itself creates new copies of the data every time it reproduces itself, thus making it an ideal archival storage system [7]. Each artificial DNA strand can hold up to 100 bits of data and preserved by making multiple copies of the DNA and inserting the original as well as the identical copies into the bacterial genome

Table 2.1: Comparison between DNA storage and Electronic Storage

Parameter	**Digital Storage Device**	**DNA Storage**
Basic Element	Transistor, Magnetic or Optical Domain	Nucleotide
Basic Element Size	$1,000/10,000nm^2$	$1mm^3$
Addressing Scheme and Organization	Discrete, Hierarchical	Sequential, Parallel Copying
Redundancy	Multi Layer	Data Replication
Capacity	Unlimited	0.1 GB / genome
Error Rate	10^4	10^-9

sequence. These copies work as back-ups of the data to counteract natural degradation. The main difference is the data stored in DNA can never be lost, but in other storage devices there is always a chance for Data loss. Table 2.1 compares DNA and Electronic Storage.

2.4 Digital Data Storage on DNA

This paper is by Raunak Laddha and Kishor Honwadkar from Smt. Kashibai Navale College of Engineering Pune.

2.4.1 Introduction

As the data increases, the current data storage technology would not be enough to store data in future as data is growing every day. Even potentially important information can get lost due lack of storage space. One of the most common cause of data loss is accidental deletion of files without backup. Every day many people lose important data because of deleting files accidently because they do not have proper backup systems. 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. In previous techniques, the data is stored in long virtual DNA molecule but encoding is done using synthetically prepared short DNA strand [8]. Short strands will allow to easily manipulate 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 [9].

2.4.2 Proposed Method

In the model proposed in this paper, ssDNA 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.

2.4.2.1 Encoding

1. Form frequency table of characters of the data.

2. Now Huffman tree of non-repeating nucleotides for encoding is generated using a set of rules.

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 1^{st} segment.

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

6. Reverse complement 2^{nd} 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.

2.4.2.2 Decoding

1. The decoding process is simply the reverse of the encoding process.

2. The 1^{st} nucleotide of DNA will tell whether the DNA is the 1^{st} or 2^{nd} segment of the pair, whether the data is reverse complemented or not and directionality of strand.

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

4. Now by using the same Huffman tree, data can convert the data into original characters.

Figure 2.6 tells us about how the DNA strands will be arranged. The green part is orientation nucleotide which will tell about the direction or orientation of DNA. The orange part is the data or information encoded which is not reverse complemented. The yellow part is the data which is reverse complemented. The blue part is the index of the strand. 2 strands of DNA always run in opposite direction and the pairs that can form a bond are specific [10].

A can a form with T and C can form a bond with G. Due to these conditions, alternate strands were reverse complemented.

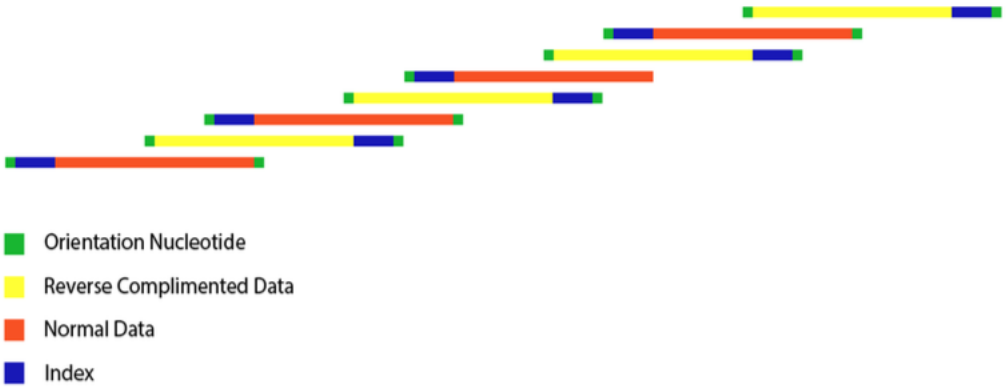

Figure 2.6: Position and Arrangement of DNA strands

2.4.3 Advantages

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.

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. It offers high loss less compression and security to data. Random access of data is also possible using this method.

Table 2.2: Comparison between DNA storage and Electronic Storage

File name	**Size in bits**	**Length after encoding**	**Total nucleotides required**
Watsoncrick.pdf	2943008	1001699	2123604
pic.png	930360	305289	647236
Dummy.txt	8192	3162	6784
IJCA paper template.docx	189944	63070	133772

2.4.4 Results

Table 2.2 shows the size of the file and total nucleotides that will be required to store the data of file using the encoding process. The file name column gives information about the file, size in bits column is size of the file in bits, length after encoding column tells about the length of data after encoding, and total nucleotides column tells about the number of nucleotides required to store the data in the required format along with a copy of data with is reverse complemented.

The number of copies can be increased by changing the offset value. The length of data in each segment can be increased by increasing the size of DNA strand. Index size can be increased or decreased to fit the need of user but the method to create index should not change to avoid repetition. In this, as there are more outgoing branches in Huffman tree compared to Huffman tree used for binary data, the data is relatively more compressed. To convert other types of files, the files need to be read in hex, binary, etc. types depending on needs and so this process is robust.

Chapter 3

Conclusion

Current data storage only lasts around 100 years, but this technology could allow the safe storage of data for millions of years. Information storage using DNA is robust for more than one hundred million years. According to the researchers, bacteria have particularly compact DNA which is passed down from generation to generation. Although mutations do occur as the data is passed from generation to generation, the rate should be slow enough to maintain the data integrity. The artificial DNA that carries the data to be preserved makes multiple copies of the DNA and inserts the original as well as identical copies into the bacterial genome sequence. The multiple copies work as backup files to counteract natural degradation of the preserved data.

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.

References

[1] Mohan s, vinodh s and jeevan f r. preventing data loss by storing information in bacterial dna. international journal of computer applications.

[2] Bonnet, j. et al. chain and conformation stability of solid-state dna: implications for room temperature storage.

[3] Erlich,h.a.,gelfand,d.sninsky,j.j.recent advances in the polymerase chain reaction.

[4] Screening framework guidance for providers of synthetic double-stranded dna federal registrar 75, 62820-62832 (2010) fr doc no: 2010-25728.

[5] George m. church, yuan gao, sriram kosuri. next- generation digital information storage in dna. science, 337, 1628 (2012).

[6] Quantitative analysis of dna orientation in stationary ac electric fields using fluorescence anisotropy seiichi suzuki, member, ieee, takeshi yamanashi, shinichitazawa, osamu kurosawa, and masao washizu, senior member, ieee 2008.

[7] C. bancroft, t. bowler, b. bloom, c. t. clelland. long- term storage of information in dna. science, 293, 1763 (2001).

[8] 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).

[9] 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).

[10] Ailenberg, m. rotstein, o. d. an improved huffman coding method for archiving text, images, and music characters in dna. biotechniques 47, 747754(2009).

Appendices

Appendix A

Additional Pictures regarding Data Storage in DNA.

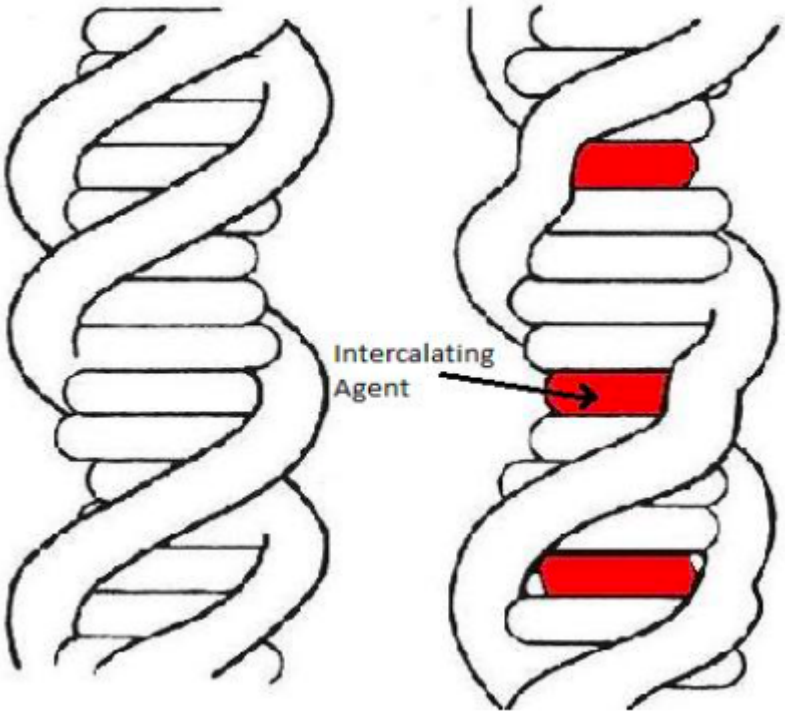

Figure A.1: Formation of Intercalating Agent in DNA

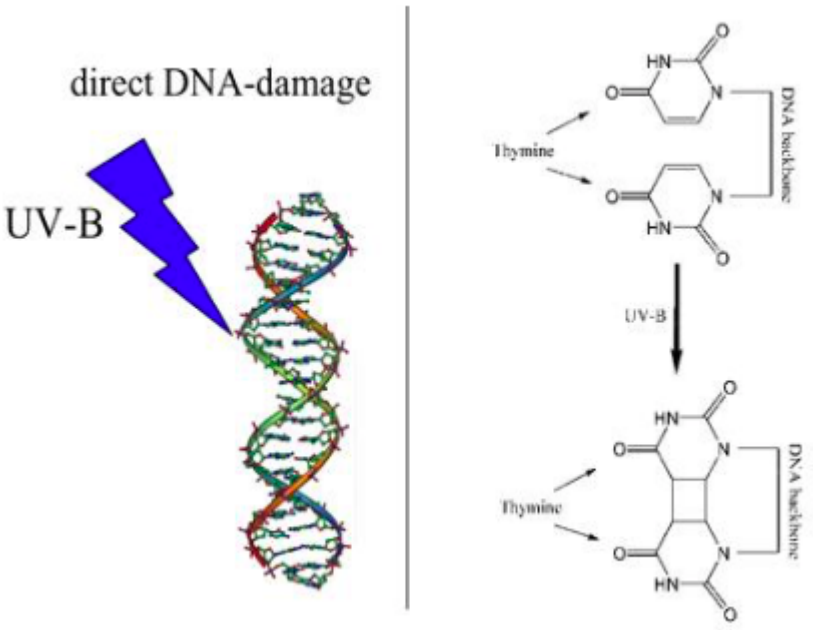

Figure A.2: DNA Responding to UV rays

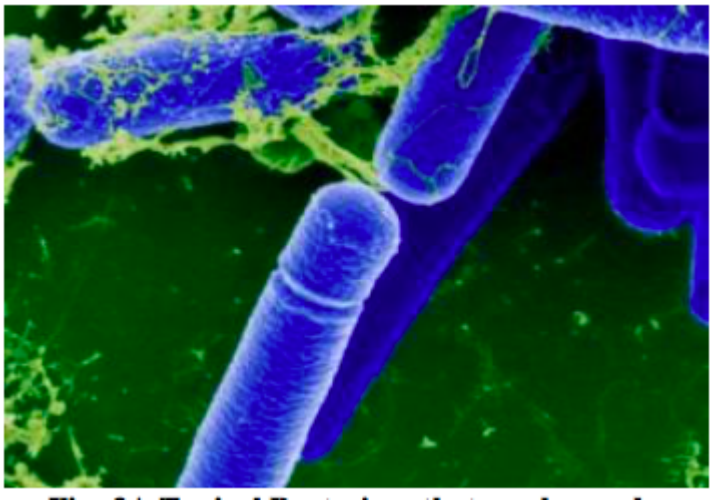

Figure A.3: Typical Bacterium that can be used

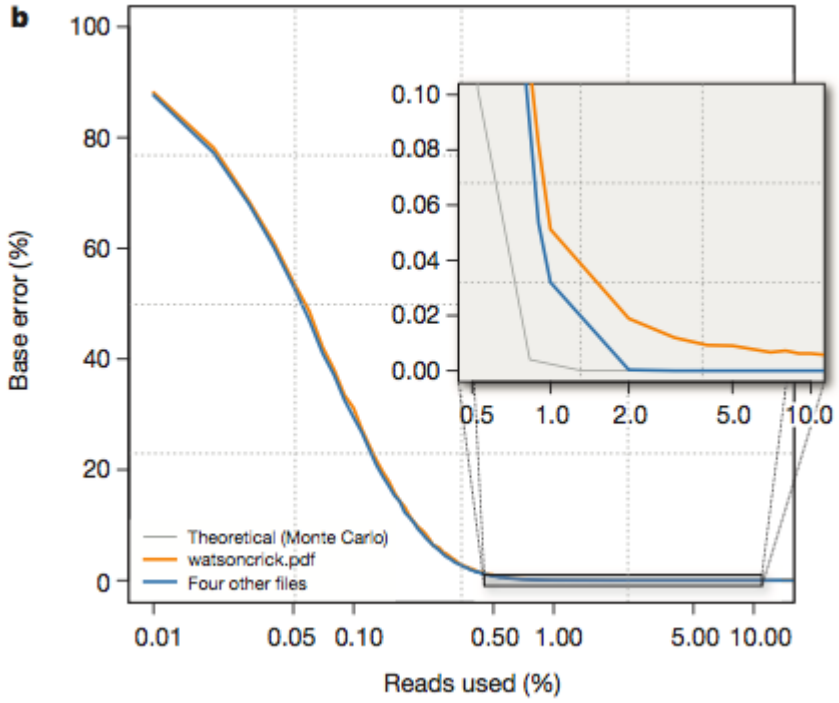

Figure A.4: Per-recovered-base error rate as a function of sequencing coverage

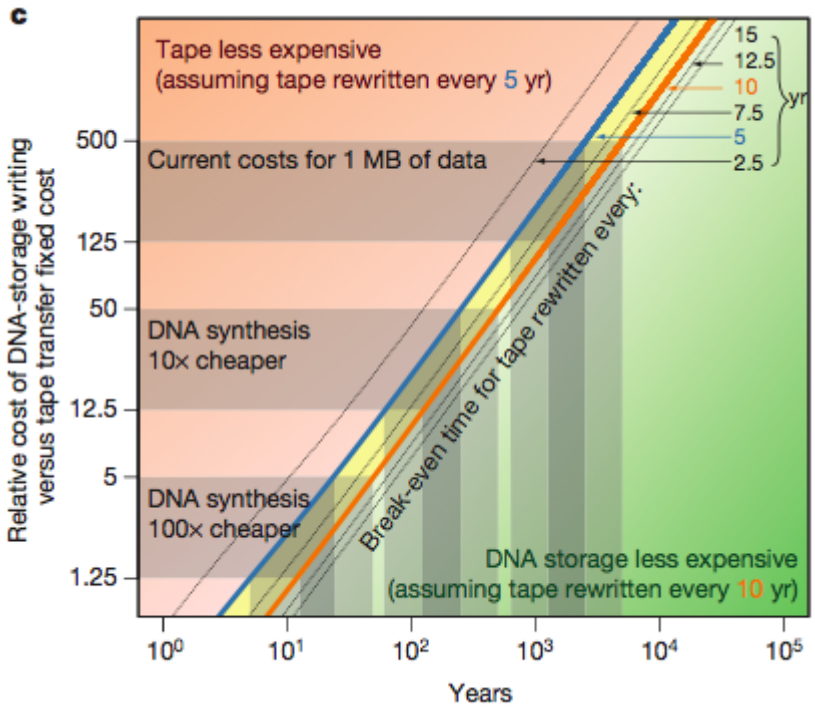

Figure A.5: Timescales for which DNA-based storage is cost-effective

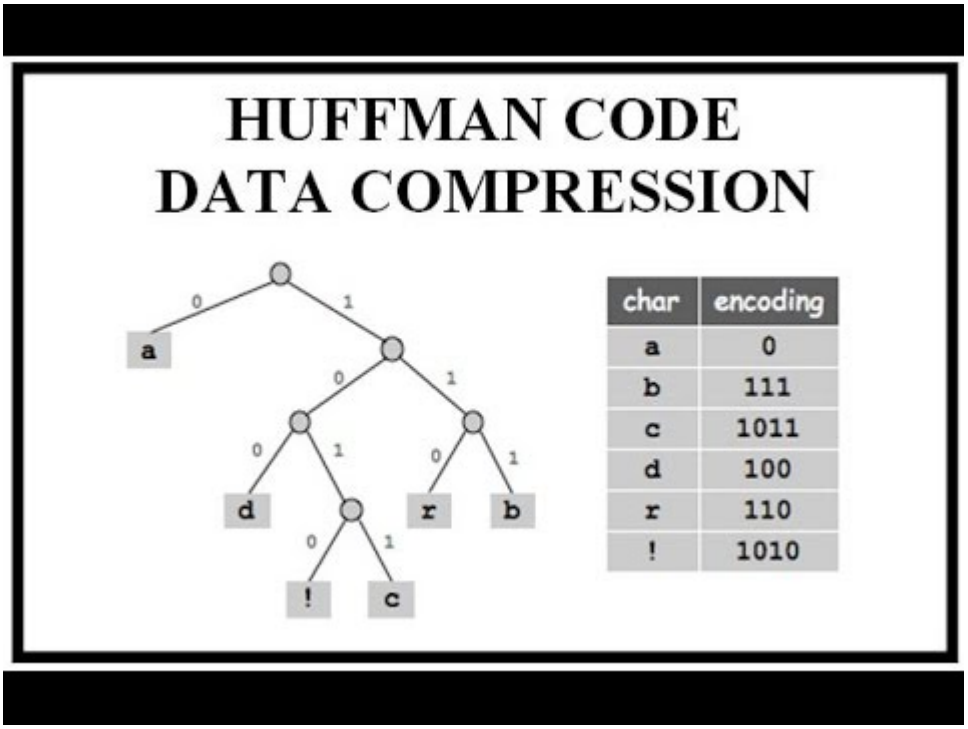

Figure A.6: Huffman Tree for DNA Encoding

Appendix B

Base paper used for preparation of Natural Computing Assignment Report.