Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 7, Number 1 (2014) pp. 1-11

http://www.ripublication.com

An Insight into an Efficient Digital Data Storage Capacity in DNA

Muhammad Rukunuddin Ghalib[1], Salunke Avinash N[2], Shruti Gupta[3], Varsha Agarwal[4]

School of Computing Science and Engineering,
VIT University Vellore, Tamil Nadu, India.
[1]ghalib.it@gmail.com [2]avinsalunke@gmail.com
[3]10julyshruti@gmail.com [4]varshaagrwal12@gmail.com

Abstract

This paper suggests the new approach for storing the digital information in DNA nucleotides. Today's technology facing the problem for storing the Big data on existing system. So DNA can be best solution to solve this problem. DNA is basic source of storing the genetic information so this paper explains the methodology to solve it. Digital file data is converted into DNA nucleotides base pairs and forming the artificial DNA of the sequence which is preserved neatly. We developed mapping standard for conversion of digital data to DNA base pairs. The idea and methodology explain in this paper can motivate the new generation of Bio Storage Technology.

Keyword- Big Data, DNA Nucleotides, artificial DNA Data storage, Bio storage Technology

I. INTRODUCTION

In present scenario data is stored on Digital Disk such as Hard Disk, Flash Drives etc. but these are less concerned about the long-term viability of your data and storage capacity. With a rapid increase in storage requirements, day by day it is increasing every year, this storage requirements are challenged to form new system which can overcome on it. This streamlining forms the combination of Biology and Computer Science so that it leads to introduce new DNA technology and the movement of data to different tiers of storage so that storage resources are optimally deployed. Several solutions are available today that can help in storage. Each of the existing approaches has its benefits and drawback, So in this paper we proposed the idea of storing Digital information in DNA which is the main source of Biological (genetic) Information

Storage in living organisms[1, 2]. The data is store on DNA in the form of nucleotides with a specific base pair. There are basically four nucleotides such as A, C, G, T. The combination of these nucleotides will form sequence of DNA strands and these DNA strands are combining using DNA Helix model to retain the biological properties of living organisms. By using the combination of these nucleotides we can form 4^4 i.e. 256 combination of possible strings comparing to Digital system which gives combination of 2^4 i.e. 16 possible combinations. Here world length is considered as 4. Hence by using this technique we can gain huge amounts of memory space for storing as well as long term viability of data [2]. In the following figure 1, detail procedure of Digital data to DNA conversion is given which explain the concept of paper. Here ”Computer” is string which converted into DNA nucleotide which has shown in flow chart.

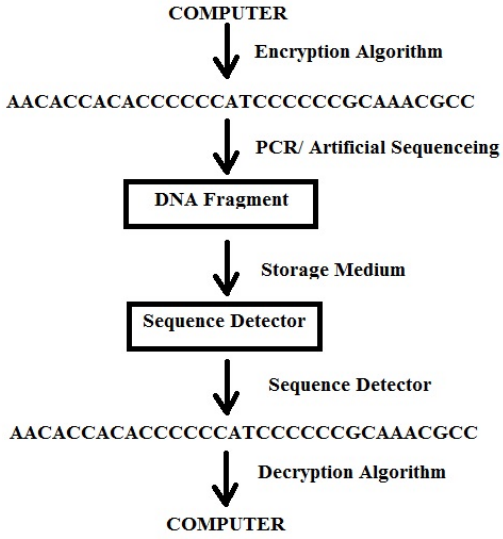


Figure 1: Steps to convert Digital data to DNA nucleotides sequence

DNA is basically a very large molecule that serves as a permanent store of the genetic information. These characteristics of DNA molecules are even more similar to the sequences of zero’s and one’s that digital computers use to represent the information [1]. DNA is an attractive target for information storage that can reliably store more information than has been handled before [3]. We encoded computer files into a DNA code the security of the encoded message was maintained by encryption method, synthesized this DNA, sequenced it and reconstructed the original files [12]. Data encoding method is performed into two phases. The first phase is to get the original text message or file which is considered for encryption [3]. Second step is introduction of mapping table, which maps the given digital keyboard characters to nucleotide sequence [4]. The decoding of message can be performed by reversing the encoding scheme [12].

II. RELATED WORK

The existing system stores the Digital Data in the digital media. If we want to store the information such as ”First Nuclear Test was taken by India in 1974 in Pokhran and the code name was Smiling Buddha” courtesy: http://en.wikipedia.org/wiki/Smiling_Buddha. To storing this digital information we have to convert into the ASCII values of it and append one extra bit to MSB (Most Significant Bit). Hence this binary form which is in 8 bit long can be store on Hard Disk. Here total 93 characters are present hence for storing this we required 93*8=744 bits which is 93 Bytes. And also the storage media which we are using such as Hard Disk are persistent of specific amount of period and with specific amount of memory storage. Drawback of this system is that after the few decades there is chance to loss of data which was store on it and also the limited memory space for storing huge amount of data [1]. Hence to avoid this we have proposed a Technique which can solve these problems and gives the memory flexibility [5].

III. PROPOSED SYSTEM

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, from which more than 99 percent of those bases are the same in all people [2]. In following figure shows DNA double helix model [18].

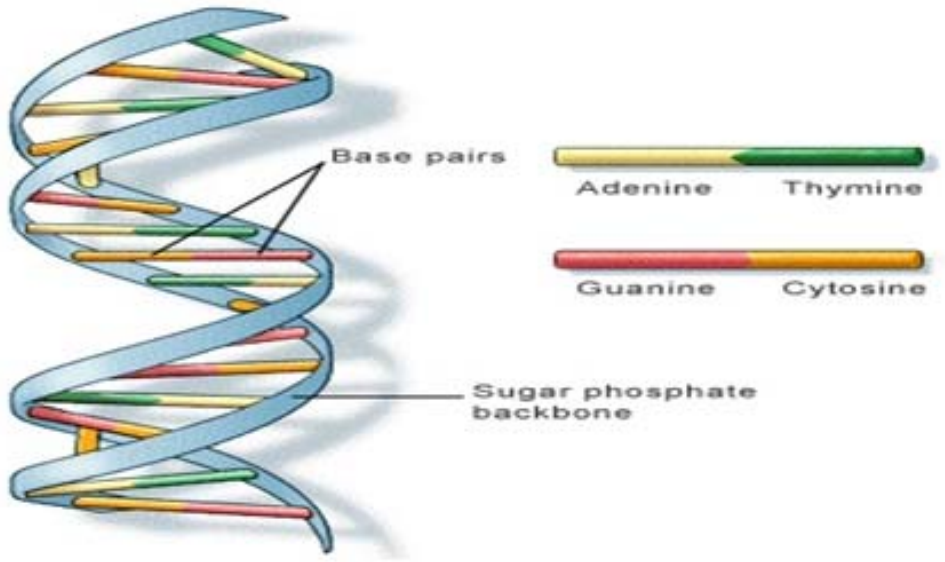


Figure 2: General DNA Helix Model (courtesy: U.S National Library)

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. Each base is also attached to a sugar molecule and a phosphate molecule. Nucleotide is a formation of a

base, sugar, and phosphate molecule. These are arranged in two long strands that form a spiral called a double helix model. The structure of the double helix model is liked a ladder, these base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder. So the main idea is we have taken the input string and this input string will be map with the DNA nucleotides. Each letter will be of four nucleotides that mean for single letter we are mapping four nucleotides base with its complementary base pair.

IV. METHODOLOGY

1. Mapping with nucleotides

Generally in digital system the data is store in the form of binary format 0 and 1. Nucleotides are represented as A, C, G and T [12]. But in reality the keyboard letter are made for human understanding and binary for Computer System. So here mapping is directly between human understanding languages to biological DNA. One letter of keyboard is mapped with four nucleotides base pairs resulting in 256 combinations as shown in mapping table 1. Reason for choosing the four nucleotides base pair is that we have consider the ASCII value table which gives the mapping of letter in decimal format which again map into binary format. Hence instead of converting decimal to binary format, we have directly mapped with DNA base pairs which reduces the complexity for designing purpose. This mapping is shown in Table 1 of digital to DNA nucleotides. After this mapping we get the DNA nucleotide sequence which is already in encrypted format so that data integrity can be obtained and mapping table provides standard for conversion digital data to DNA and vice versa. This mapping can be shown mathematically as follows which is the basic steps in algorithm used for conversion.

Basic initialization:

K= { (A, B, …, Z), (a, b, …, z), (0, 1,..9), (~, !, …/)} ←set of keyboard letters
D= {A, C, G, T} ← set of DNA nucleotides
M={x: x ↔ y || x € D and y € K} ← Digital to DNA mapping function
Rm={y: y ↔ x || x € D and y € K} ← DNA to digital mapping

Using the above mathematical relationship set K is converted into set D using relationship M in the following equation 1.

$$\mathrm{F}(M) = \mathrm{K} \rightarrow D \tag{1}$$

And for reversing i.e. conversion of DNA to digital set D is converted into K using relationship R_m as given by the equation 2 below.

$$\mathrm{F}(R_m) = D \rightarrow \mathrm{K} \tag{2}$$

Algorithm:

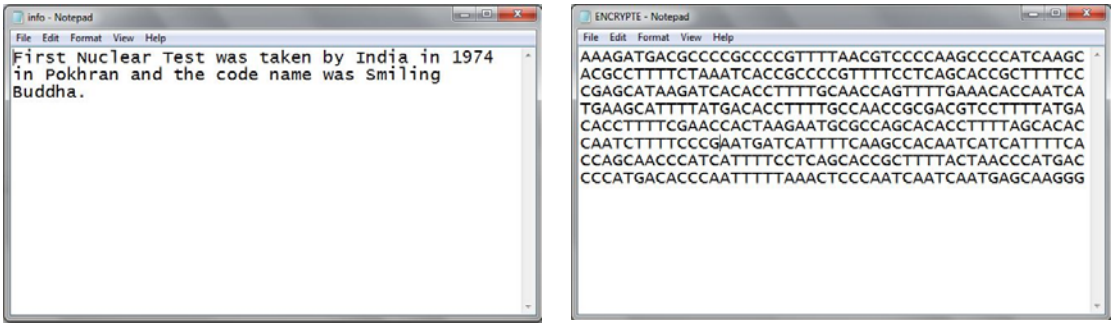
1. Take string as set of K.
2. Map string K with set D by following the mapping equivalence given in table 1 using mapping function M.
3. Formation and storage of the DNA sequence obtained from the step 2.

4. Preservation of DNA drop on Bio chips in laboratories.
5. Reverse mapping from set D to string K from bio chip sequences and read by sequence detector algorithm.
6. Reverse mapping is done by R_m function by following the mapping equivalence given in table 1.
7. Successfully retrieve the digital string which was store on DNA.
8. Stop the process.

Table 1: Digital to DNA nucleotide mapping tabulation

Character	Nucleotide Base	Character	Nucleotide Base	Character	Nucleotide base
A	AAAA	G	CAAT	“	TGCC
B	AAAC	H	AATG	#	GCCT
C	AACA	I	ATGA	$	CCAT
D	ACAA	J	TGAA	%	CCTA
E	CAAA	K	TAAG	‘	CATC
F	AAAG	L	CCCC	(	CTAC
G	AAGA	M	ACCC	)	ATCC
H	AGAA	N	CACC	*	TACC
I	GAAA	O	CCAC	+	TCCA
J	AAAT	P	CCCA	,	ACCT
K	AATA	Q	GCCC	-	GGGG
L	ATAA	R	CGCC	.	AGGG
M	TAAA	S	CCGC	/	GAGG
N	AACG	T	CCCG	&	GGAG
O	ACGA	U	TCCC	@	GGGA
P	CGAA	V	CTCC	:	TGGG
Q	GAAC	W	CCTC	;	GTGG
R	AACT	X	CCCT	<	GGTG
S	ACTA	Y	CCAG	>	GGGT
T	CTAA	Z	CAGC	=	CGGG
U	TAAC	0	AGCC	?	GCGG
V	AAGT	1	GCCA	[	GGCG
W	AGTA	2	CCGT	]	GGGC
X	GTAA	3	CGTC	`	TTTT
Y	GAAT	4	GTCC	_	ATTT
Z	AAGC	5	TCCG	~	TATT
A	AGCA	6	CCGA	{	TTAT
B	GCAA	7	CGAC	}	TTTA
C	CAAG	8	GACC	|	GTTT
D	AATC	9	ACCG	^	TGTT
E	ATCA	Space	CCGT	\	TTGT
F	TCAA	!	CTGC	DEL	TTTG

The mapping of digital data with nucleotides can be done by using software which uses above mapping table and convert the string into appropriate DNA form. For input to the software we have used info.txt which contain the string of information ”First Nuclear Test was taken by India in 1974 in Pokhran and the code name was Smiling Buddha”. This information and converted DNA information info.txt and encrypt.txt files as shown in figure 3.



(a) (b)

Figure 3: The files used as source for string input and the output file after encryption (a) Source file- info.txt file (b) Encrypted source file-encrypt.txt

The software for conversion is shown in figure 4. This software takes input as file and gives the output in the form of file. In this paper we have taken files as input and file can contain any information, it may be data, audio, video, images etc. Figure 4 shows both operation of conversion of digital info into DNA nucleotides and vice-versa.

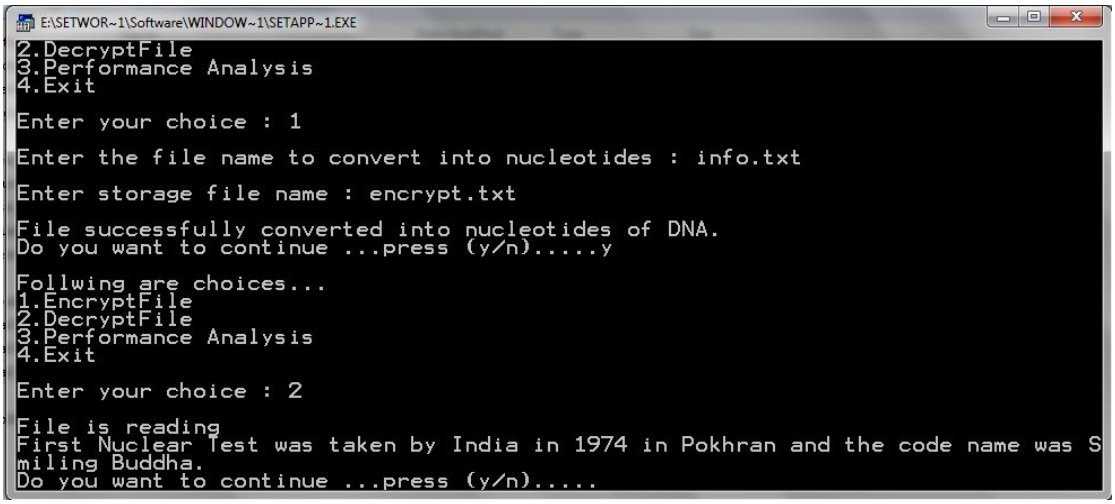


Figure 4: Conversion of digital data into DNA nucleotides and reversers using SETAPP software.

2. Evaluation of mapping

After mapping the string into nucleotide base pairing we can evaluate the performance with respect to memory size. Each character will be encoded into 4 base pair and in our string we have 93 characters so that 93*4 = (372+4) bp=376 bp (bp ➔base pair) nucleotides sequence will be formed. And this sequence is as one strand, for forming the double helix model of DNA we required two DNA strands which are complementary to each other's [20]. The double helix model made up of two DNA strands which have shown above. So when we encrypted the string into nucleotides and storing in one file called "encrypted.txt" and at the same time we have form another file which is complementary for this file named as"complement.txt". To forming the double helix model we pair the nucleotides simultaneously from "encrypted.txt" and from"complement.txt"[14]. By implementing this methodology we have achieved error detection mechanism while conversion. The sequence we obtained by forming double helix model will be stored on DNA by using chemical synthesized DNA and PCR machine.

3. Synthesized DNA

The sequence we obtained by DNA synthesis is a natural or artificial creation of Deoxyribonucleic acid (DNA) molecules [6]. In nature such molecule are created by all living cells through the process of DNA replication with replication initiator proteins splitting the existing DNA of the cell and making a copy of each split strand, with the copied strands then being join together with their template strand into a new DNA molecules [10]. Various means also exist to artificially stimulate the replication of naturally formation of DNA or creation of artificial gene sequences.

A) PCR (Polymerase Chain Reaction): A polymerase chain reaction is a form of DNA synthesis from enzymatic reactions [15], using cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA [7].

B) Artificial Gene Synthesis: Artificial gene synthesis is the process of synthesizing a gene in vitro without the need for initial template DNA samples [23]. Artificial gene synthesis is a method in synthetic biology that is used to create artificial genes in the laboratory. Based on solid-phase DNA synthesis, it differs from molecular cloning and polymerase chain reaction (PCR) in that the user does not have to begin with pre-existing DNA sequences. Therefore, it is possible to make a completely synthetic double-stranded DNA molecule with no apparent limits on either nucleotide sequence or size

C) Oligonucleotide Synthesis: Oligonucleotide synthesis is the chemical synthesis of relatively short fragments of nucleic acids with defined chemical structure (sequence) [19]. The technique is extremely useful access to custom-made oligonucleotides of the desired sequence [25]. Whereas enzymes synthesize DNA and RNA in a 5' to 3' direction, chemical oligonucleotide synthesis is carried out in the opposite, 3' to 5' direction. The process is implemented as solid-phase synthesis using phosphoramidite method and phosphoramidite building blocks derived from protected 2'-deoxynucleosides (dA, dC, dG, and T), ribonucleosides (A, C, G, and U), or chemically modified nucleosides.

V. STORAGE/ PRESERVING THE DNA FRAGMENT

After generating the DNA sequence with the above techniques we have to preserve these DNA fragments as given in the literature of Anchordoquy, T.J. and Molina, M.C, (2007) [16]. The drop of this DNA fragments can be store on bio chips. This bio chips can be store in preservation laboratories [8, 9]. Also we can store the dehydrated DNA at room temperature which is completely protected from water and oxygen [21]. Here the preservation is mainly referred to maintenance of chemical and physical integrity of the DNA molecule [24].

Following are the DNA storage strategies [13]

- Short-term storage (weeks) at 4°C in Tris-EDTA
- Medium-term storage (months) at –80°C in Tris-EDTA
- Long-term storage (years) at as –80°C as a precipitate under ethanol [22]
- Long-terms storage (decades) at –164°C or dried

VI. SEQUENCE DETECTOR

A DNA sequencer is a scientific instrument used to automate the DNA sequencing process. Given a sample of DNA, a DNA sequencer is used to determine the order of the four bases. The order of the DNA bases is reported as a text string, called a read [11]. Some DNA sequencers can be also considered optical instruments as they analyse light signals originating from fluorochromes attached to nucleotides. For detection of this sequence we are using existing algorithms such as shotgun sequencing. Sequence detector has detected the sequence from synthesize DNA which is input to decryption algorithm. This decryption algorithm again form the Original string ”First Nuclear Test was taken by India in 1978 in Pokhran and the code name was Smiling Buddha”.

VII. EXPERIMENTAL RESULT AND ANALYSIS

For performance analysis, consider the same string ”First Nuclear Test was taken by India in 1974 in Pokharan and the code name was Laughing Buddha”. After converting this string calculate the performance metric of the two system such as Digital Data storage and DNA storage. For these both systems we have taken the information in the form of file and evaluating the total storage size on Hard Disk and DNA. The figure 6 shows details of performance analysis by using the same software. In that digital file takes 93 characters which took 93 bytes (as each character takes 8 bits). While DNA base pare takes 376 base pairs and for storing the same base pairs are used.

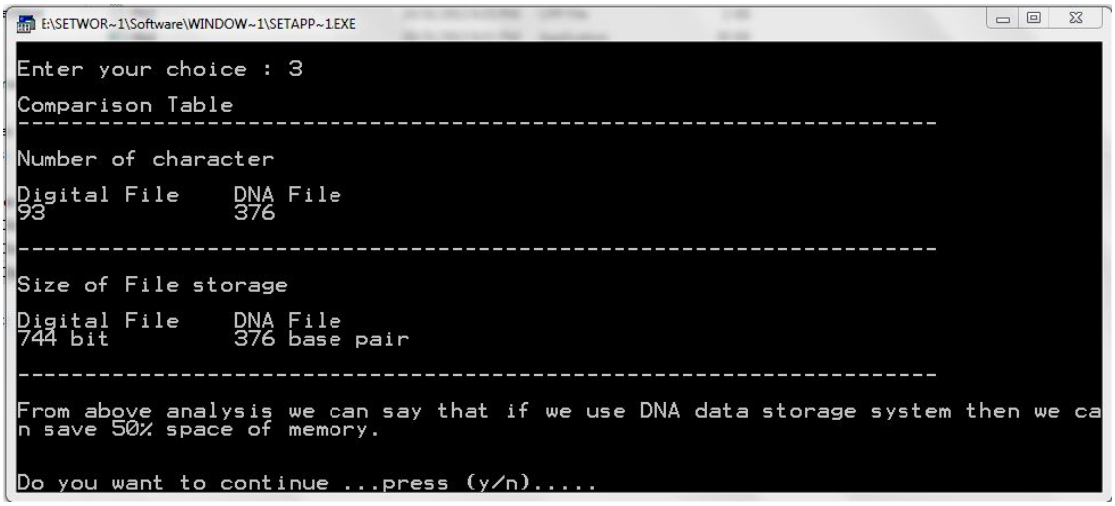


Figure 6: Performance analysis of Digital string storage to DNA based storage system with respect to data size.

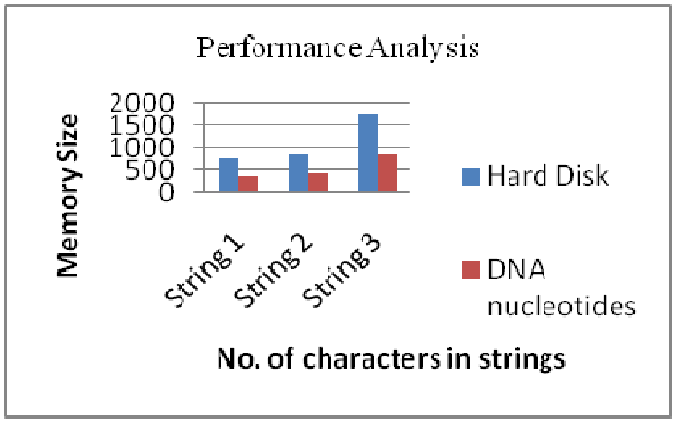


Figure 7: Performance Analysis

The above figure 7 is a graph showing number of character with respect to memory size. From above analysis we have proved that DNA storage can be best option to store the information instead of digital data. DNA storage system requires 50% less space than digital data storage system of the same size. The String 1, String 2 and String 3 are stored on Hard Disk which requires space of 744 bits, 880 bits, and 1760 bits respectively. Here String 1 is actual string which is used for conversion and String 2 and String 3 are similar synthetic strings taken for analysis purpose only. And same strings when store in DNA, those requires space of 376bp, 440bp, 880bp. The histogram shown above for data storage as Hard disk with respect to DNA storage concludes that DNA nucleotides utilize better space than Hard Disks.

VIII. CONCLUSION AND FUTURE WORK

With the basic functionality of DNA to represent genetic information, DNA can be used as the best source for digital information storage. With this DNA storage technology we have optimized up to 50% space compare to digital disks. This technology stores the data without corruption for longest duration of time. This feature will help for data storage in big organizations such NASA, ISRO, BARC, DRDO and data warehouse etc. Because DNA is biological source so there is no wastage like electronic waste. Hence this paper motivates the formation of New Biological Data Storage Technology which can make reform in Computer history.

In future, we can develop the interface between the PCR and the real computer systems to reduce the time complexity of data conversion. We also plan to design a efficient biological chip architecture for the storage of encrypted DNA sequences generated from our proposed work.

REFERENCE

[1] Niok Goldman, (2013) ”Towards Practical, High-Capacity, Low Maintenance Information Storage In Synthesized DNA”, Nature 495, DOI:10.1038/nature 11875.
[2] Georage M. Church, Yuan Gao, Sriram Kosuri (2012), ”Next Generation Digital Information Storage In DNA”, DOI:10.1126/science.1226355.
[3] Lin Edwards (2013), ”DNA used to encode a book and other digital information”, phys.org
[4] Jonathan Keith (2013),”DNA data storage: 100 million hours of HD video in every cup”, phys.org/news/
[5] John Bohannon (2012), ”DNA: The Ultimate Hard Drive”, Science Now.
[6] Tim Barribean (2013), ”Artificial Life Synthetic DNA that can self replicate”, Science 5543843.
[7] Erlich, H.A. Gelfand, D. Sninsky, J.J. (1991), ”Recent Advances In the Polymerase Chain Reaction.”, Science 252, 1643-1643-1651.
[8] Bonnet, J.et al (2010), ”Chain and conformation stability of solid-state DNA: implication for room temperature storage”, Nucleic Acids Reg.: 38, 1531-1546.
[9] Achordoquy, T.J & Molina, M.C, (2001)”Preservation of DNA”, Cell Preservation Technology 19, 247-250.
[10] Lutz JF, Ouchi M, Linu DR, Sawamoto M, (2013), ”DNA Sequencing Controlled Polymers”, DOI:10.1126/science.1238149
[11] The University Of Michigan, ”DNA Sequencing Core”
[12] Rahul Vishwakarma, (2012) ”High Density Data Storage In DNA using an Efficient Message Encoding Scheme”, IJITCS Vol.2, No.2
[13] Oxford Gene Technology, (2011), ”DNA Storage and Quality”. http://www.ogt.co.uk/resources/literature/ 403_dna_storage_and_quality.
[14] Guangzhao Cui, ”An Encryption Scheme Using DNA Technology”, Zhengzhou University Of Light Industry, Zhengzhou.
[15] Junghuei Chen and Yuzhen Wang, ”The Ultra High Density Storage Of Non-biological Information In A Memory Composed OF DNA Molecules”, University of Delaware, Newark, Delaware.
[16] Anchordoquy, T.J. and Molina, M.C, (2007) ”Cell Preservation Technology”.

[17] http://en.wikipedia.org/wiki/Smiling_Buddha
[18] Hayam Mousa, Kamal Moustafa, Waiel Abdel, (2011) ”Data Hiding Based On Contrast Mapping Using DNA Medium”, The International Arab Journal Information Technology.
[19] LeProust, E. M. et al. (2010), ”Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process”. Nucleic Acids Res. 38, pages 2522–2540.
[20] Watson, J. D. & Crick, F. H. C., (1953) Molecular structure of nucleic acids. Nature 171, 737–738 (1953)
[21] Bonnet, J. et al. Chain and conformation stability of solid-state DNA: implications for room temperature storage. Nucleic Acids Res. 38, 1531–1546 (2010)
[22] Cox, J. P. L. Long-term data storage in DNA. Trends Biotechnol. 19, 247–250 (2001)
[23] http://en.wikipedia.org/wiki/Artificial_gene_synthesis
[24] Anchordoquy, T. J. & Molina, M. C. Preservation of DNA. Cell Preserv. Technol. 5, 180–188 (2007)
[25] Tian J, Gong H, Sheng N et al. (December 2004).”Accurate multiplex gene synthesis from programmable DNA microchips". Nature 432 (7020): 1050–4

Author’s Profile

Muhammad Rukunuddin Ghalib
He has B.Tech (IT) and M.E. (CC) from Anna University, Chennai and right now on the verge of completing his PhD in CSE from Anna University, Chennai. Presently he is an Assistant Professor (Senior) in SCSE, VIT University, Vellore from last 6 years. His research interest includes Data Mining in Bioinformatics especially in Microarray Data Analysis, Neural networks and fuzzy logics algorithm design, Protein Structure prediction and Medical Image Processing. He has several peer-reviewed national and international conferences and journals publications.

Salunke Avinash N
He is pursuing his M.Tech in CSE from VIT University currently. His areas of interest in research includes DNA storage technology, Bioinformatics, Neural Networks, Parallel algorithms. He has published few papers in national conferences earlier.

Shruthi Gupta
She is pursuing his M.Tech in CSE from VIT University currently. Her areas of interest in research include Neural Networks and fuzzy systems, Bioinformatics. She also has few papers in national conferences earlier.

Varsha Agarwal
She is pursuing his M.Tech in CSE from VIT University currently. Her areas of interest in research include DNA storage technology, Bioinformatics, Neural Networks, Parallel algorithms. She has already published few papers in national conferences.