Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

A Detailed Study of DNA Based Data Storage Techniques

Himani Rajput[1], Sonakshi Khosla[2], Riyanshi Mittal[3], Vasudha Arora[4],
Shaveta Malik[5], Dr. Prateek Jain[6]
Department of Computer Science & Engineering[1,2,3,4,5]
Manav Rachna International Institute of Research and Studies, Faridabad, India – 121001[1,2,3,4,5]
Accendere Knowledge Management Services, New Delhi[6]

Abstract: **A DNA storage system is made up of a DNA synthesizer that is used to encode the data to be stored in the DNA, a storage container and a sequencer to read DNA sequences and to convert them back to digital data. Characteristics of DNA like less use of energy, less error rate, a long shelf life, possibilities of many copies of data possibly to be obtained, high density (large amount of data stored in small place) and short strands (for easy manipulation of data), and High memory space (e.g. 1 gram of dry DNA can store about 455 exabytes of data) make DNA very compatible and reliable. DNA can retain data for much longer period of time (like around 100 years), whereas the existing storage devices like hard disc and flash memory can retain for very short time (comparatively, around 10 years). This paper provides an overview of the current approaches used for DNA-based storage system. The basis of these works is the building up as well as the structure of the sequences over distinct alphabets. These alphabets do not have pre-defined pattern of addresses, contain base content that is balanced, and also provide other substring constraints. These arrangements or techniques accommodate the gathered signals to the DNA intermediately. Adapting the stored signals to the DNA medium hence reduces the inherent number of errors in the system.**

1. INTRODUCTION

Two architectures for DNA based storage have recently been discussed. The density of first architecture was found to be 700 TB(terabytes) per gram whereas the density of second architecture was 2 PB(petabytes) per gram. This was comparatively better than first approach because second architecture made use of coding schemes such as Huffman coding, single parity check coding and differential coding. On August 16, 2012, the Science Journal prepared and issued work by George Church and his co-workers at Harvard University. They converted a draft of a book, coded in html language that included around 53,425 words, 11 images of jpg format and 1 JavaScript program into a 5.26-megabyte set of data in binary form. These bits were encoded onto oligos which were incorporated by ink-jet engraved DNA microcircuits. They were able to recover all the data blocks with not more than 10-bit faults out of 5.26 million bits. They mapped the bits one-to-one with the bases. This approach faced certain issues, such as, long repetitions of the same base was created, which made the DNA sequencing error prone. In 2013, another approach was suggested in an article by European Bioinformatics Institute (EBI). The information which they used consisted of all 154 Shakespeare's sonnets, a 26 second audio clip and much more. When the files were reproduced, an accuracy of 99.99% to 100% was measured [7]. To safeguard the data from getting lost, the researchers used an error-correcting encoding scheme. In order to avoid errors, each data region was repeated 4 times with 2 strands constructed backwards. In February 2015, in an article by researchers from ETH Zurich, it was predicted that error free recovery of data was possible till 1 million years if stored at 18° C and for 2000 years if stored at 10° C. All these methods share a disadvantage that the whole strand needs to be sequenced if we want to retrieve just one set of data out of entire encoded data set. Digital data refers to a type of the data which is represented using 0s and 1s that can be easily understood and used by other technologies. There can be other techniques of representing the data also. It may store complex audios, videos, text in form of 0s and 1s. Despite many improvements in the traditional techniques to record data, there is still so much need of high volume, non-volatile and durable recording media. DNA itself helps to implement non-volatile recoding media of outstanding results and extremely high storage capacity as an example, a human cell, with a mass of about 3 grams, can hold 6.4 GB of information of DNA encoding. One of the approach can be raised with a density of about 2PB/gram. The success to this technique was due to three coding schemes –differential coding, Huffman coding and single parity check coding.

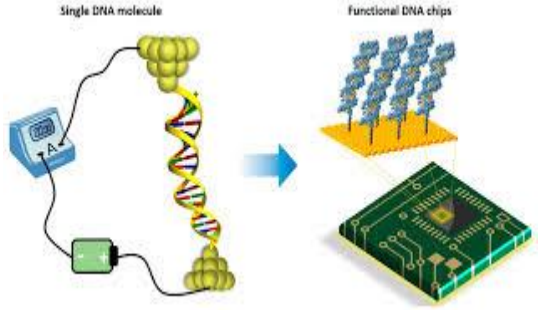


Fig 1- Single DNA molecule to functional DNA chips [6]

All the digital data storing techniques in DNA will have some drawbacks, including absence of limited approach to data.

2. DNA SEQUENCE SYNTHESIS

All the previously used techniques to store digital data have significant figure of limitations, containing the absence of limited approach to data – i.e. the whole sequence has to be reconstructed in order to read a single strand – and the rewriting mechanisms are not available. Shifting the data from a ROM to RAM, rewritable memory needs big changes in the application of the DNA storing system, because a unique addresses is to be added to composing storage in DNA structures that will not head to fallacious

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

cross-adulteration with the encoded data that the DNA structures contain ; avoiding extending DNA structures to increase the analysis and consecutive coalescence, leading to prevention of efficient rewriting; ensuring less synthesis (writing) and numbering (reading) fault estimate of the DNA structures.

To solve them and different conflicts, a hybrid design with DNA re-writable capacity and random-access abilities has been proposed. The fresh DNA-based repository system incorporates a huge amount of coding highlights, such as constrained coding, which guarantees that DNA designs are going to avoid sequencing patterns; are maintained a strategic distance from; prefix synchronized coding, with the goal that pieces of DNA might be precisely accessed without disturbing different pieces in the DNA puddle; and LDPC (low-density parity-check) coding as traditionally collected repetition battling rewriting mistakes.

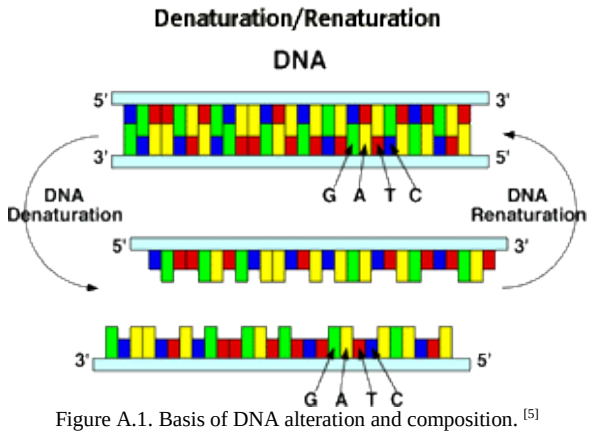


Figure A.1. Basis of DNA alteration and composition. [5]

De novo [5] DNA synthesis helps to create DNA sequences with the already existing templates. Currently, the cost is high and throughput is small of de novo composition of the construction bars means the main disadvantage which come in the way of large scale implementation of DNA synthesis systems, for example oligo synthesis via phosphoramidite column-based method, which is not used widely because of the high cost. So, there is a need to develop more robust and cheaper systems for digital data storage in DNA.

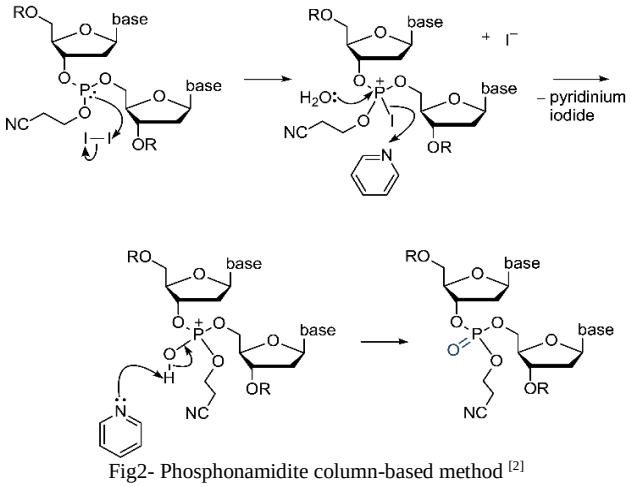


Fig2- Phosphonamidite column-based method [2]

There are other synthesis technologies like one of the very successful microarray-based synthesis methods in which around half a million of oligos can be taken care of per one microarray, which needs less of the reagent for the process to happen [3]. For DNA synthesizing projects on large scales, the price of this technique for working is roughly $0.001 per muton. Alike the phosphoramidite column-established coalescence, the dimension of microarray composited oligos generally not more than 210 nt.

Even though, oligos combined by the microarray strategy usually have high mistake rates than the ones created by the phosphoramidite column techniques. In any case, microarrays are the favored for integrating redid DNA-chips or for the blend of qualities. Different strategies are also being developed to find a system that could bring the high-cost and high-precision and minimal effort required. They will diminish the restrictions of the present strategies. For understanding the fundamental standards of DNA-supported digital retention and the cons that should be decreased during the writing computing, we first investigation distinctive DNA combination techniques from muton to the bigger DNA atoms. At that point the current systems that mean to enhance the quality and dependability of the sequences are discussed.

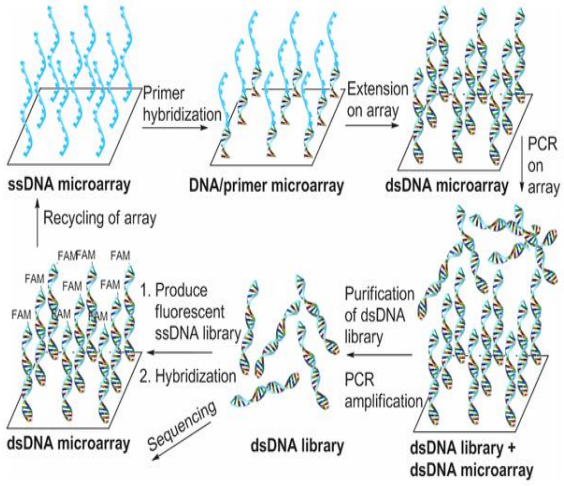


Fig 3- Microarray way of synthesizing oligos [4]

2.1 Oligo-Synthesis Platforms:

2.1.1 *Column- based Oligo Synthesis:* The officially used phosphoramidite oligonucleotide combination works through an expansion that is performed stepwise of mutons to the developing series that gets de-enacted on a strong help [4] (Fig 3). Every extra round is done in four steps: 1) un-blocking; 2) connecting or condensation 3) topping and 4) nitrification.

Towards beginning of the synthesis process, the main nucleotide, which is connected to a sturdy substrate, is completely secured at each active site. Along these lines, to have a conceivable reaction and merge another nucleotide, it is essential to move out the dimethoxy trityl (DMT), the shielding bunch from the 5'- end by including of an acidic plan. The expulsion of the DMT amass creates a receptive 5'- OH gathering (De-blocking step). Finally, a coupling

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

step is performed by methods for buildup of crisply initiated DMT-secured nucleotide and the unprotected 5'- OH gathering of the substrate-bound making oligo strand through the plan of a phosphite triester interface. (Coupling or Condensation step). After this progression, some unprotected 5'- OH packs may in any case exist and respond in later times of expansion of nucleotides affecting oligos with cancellation and enormous evacuation bungles. To relieve this issue, a topping response is performed by acetylation of the dormant nucleotides (Capping step). Finally, the insecure phosphite triester linkage is oxidized to a steady phosphate linkage utilizing an iodine arrangement (Oxidation step). The cycle is over and over iterated to get an oligonucleotide of the required succession structure. Preceding the fulfillment of the combination, the oligonucleotide game plan is unprotected, and dis-joined from the help to secure a totally utilitarian unit.

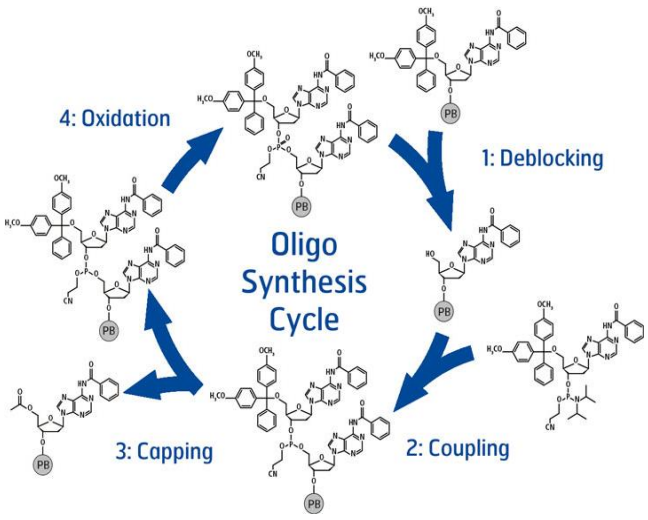


Fig 4-Main Steps of Column-Based Oligo Synthesis procedure [3]

2.1.2 *Array-based Oligo Synthesis:* One of the methods for array-based synthesis, like, the method innovated by Agilent uses Inkjet-based printing, in which with high accuracy, picolitres of each, absorbed nucleotide and activator can be discovered and settle at specific sites on an array. This ink-jet method alleviates the requirement of using photolithography masks. Both solid-phase and microarray techniques show varied challenges that are required to be overcome to reduce errors and elevate throughout. Side reactions such as depurination and reaction incapability during the addition of nucleotides reduces the wanted yield and prompts errors in the sequence especially in case of long oligo strands. [8] Notably, these processing problems introduce substitution, insertion and deletion errors. Consequently, a purification step is needed to recognize and remove the incorrect sequences. This process is very costly as the chromatography gel needed is costs huge amount. Yet, by reforming chemical reaction and conditions the precision can be increased.

2.1.3 *Complex Strands and Genetic Synthesis:* Ordinarily, for generation of DNA sections of a couple of hundred nucleotides' length, a gathering of shorter length oligostrands is consolidated either by utilizing ligation-based or polymerase-based responses. Ligation-construct techniques by and large depend in light of thermostable DNA ligases that ligate phosphorylated covered oligos in exceptionally strict conditions. In polymerase-based strategies (Polymerase cycling gathering - PCA) oligos with covered locales deliver continuously longer twofold strand arrangements. In the wake of collecting, combined successions are to be PCR opened up, cloned, and confirmed, subsequently expanding the generation cost. In spite of the fact that the cost of oligonucleotides is decreased by utilizing microarray combination, two noteworthy difficulties still hamper its utilization. Initially, a large number of oligonucleotides can be formed on a solitary microarray, however each oligo is produced in little amounts. Next, the oligo strands are divided from the cluster at the same time as a huge heterogeneous pool that in the long run prompts challenges in sequence assembly and cross-hybridization. A few strategies have been recently developed, to mitigate these issues. For instance, PCR amplification elevates the concentration of the oligos before assembling that joined with hybridization selection, lessens the incorporation of oligonucleotides containing unwanted synthesis errors. An altered approach, in view of hybridization selection set in the assembly process and combined with the recovery of oligo design and assembly conditions was reported. Still, huge pools of oligos (>10000) increase the difficulties in sequence assembly.

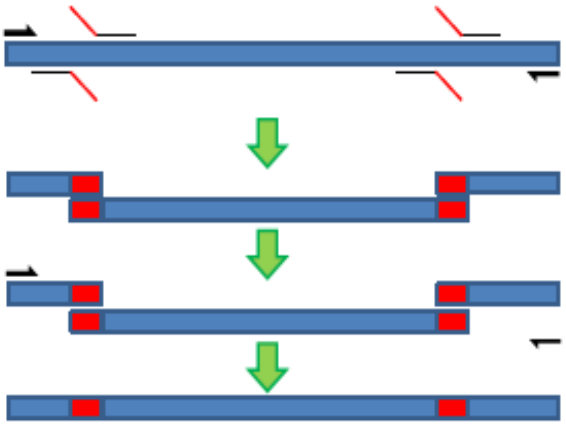


Fig 5- Rewriting (Deletion and Insertion). [1]

2.1.4 *Error-Detection:* Regardless of explained biochemical error removal procedures present, certain leftover errors tend to stay in the incorporated pool and more errors emerge amid the get assembling stage. Huge numbers of the present error removal systems depend just on DNA pattern reorganization proteins. Denaturation and rehybridization steps prompt double stranded DNA with mismatch between wrong bases and the similar right bases.

The disrupted sites are seen and cut by mismatch acknowledgment proteins. MutS is a protein that ties unpaired bases and little DNA loops that stick out from the double helix structure. After denaturation and re-hybridization, it distinguishes and, binds the mismatched areas which are later removed by gel electrophoresis. The error rate is decreased to 1 nucleotide per 10 Kb utilizing this method.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

The correct sequences are recuperated by gel electrophoresis. At last, it is observed that even single substitution mistakes in the amalgamation procedure might be risky for applications in biological and medical research. In any case, this is not the case with DNA-based storage frameworks, in which the DNA strands are used as storage system which may have a trivial error rate. Synthesis errors can be effectively battled by the presentation of precisely outlined parity checks of the data strings.

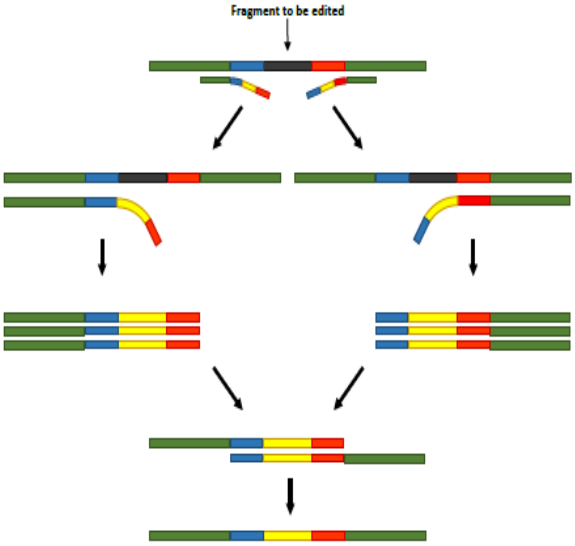


Fig 6-. Rewriting (Deletion and Insertion). [1]

3. DNA SEQUENCING

The inspiration driving DNA sequencing is to read the DNA content, to decide the correct nucleotides and their order in an atom. This information is important for creating nature-propelled computational platforms. Sanger et al. were the first ones to create sequencing procedures to sequence DNA in light of chain termination.

This strategy, normally addressed as Sanger sequencing, has been generally utilized for a long time is as yet being utilized as a part of different research centers. Nevertheless, in the previous decade, the improvement of quicker, less expensive, and higher-throughput sequencing innovations has humongous extension of the scope of genomic studies. Generally, the NGS (next generation sequencing) advancements have a few noteworthy differences compared with Sanger sequencing. To start with, electrophoresis isn't required for reading the sequencing output which would can now be directly detected. Second, more straightforward library arrangements which don't utilize DNA clones have turned into a vital piece of sequencing work process. [4]

Third, humongous number of sequencing reactions are produced in parallel with a high throughput. A show of the critical NGS innovation is the cost reduction, and the cost continues dropping after some months due to new improvements in sequencing techniques.

However, the weakness of NGS advances is data quality. The read lengths are shorter, and the error rates are higher. The primary NGS stage was achieved by "454 Life Sciences". Albeit "454 platforms" were closed down in 2016, they have made critical commitments to NGS technology development.

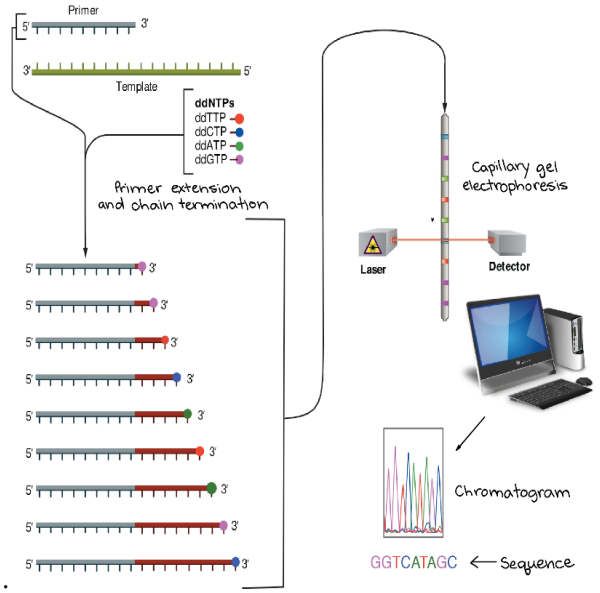


Fig 7-. Major series of the Sanger sequencing protocol. [2]

4. CHURCH GAO-KOSURI IMPLEMENTATION

A recorded DNA-based capacity design was specified of a huge scale in the original work of Church et. Amidst the given technique, client information was changed over into a DNA sequence through a pattern by pattern mapping, which encodes every bit of data 0 into A or C, and every 1 into T or G. What basis will be picked up out of the two for changing a specific bit is chosen by picking a base randomly ensuring that it prevents homopolymer strands of dimension more than 3. To exhibit the use and execution of their technique, the respective writers made a PHP document of 5.27 MB in DNA. This consisted of 53.5 k words (approximately), 11 JPEG pictures plus a Java language Scripted file. [2] For removing the requirement for long human-made DNA strings which are difficult to arrange, the record was changed over into 54, 898 pieces of length 159 oligonucleotides. The oligonucleotide library was integrated utilizing the Inkjet printers, and of high-fidelity DNA microchips. Just errors of 10 bit were seen inside the 0.5 crores encoded bits, i.e., the revealed framework fault percent was little.

EXAMPLE:

In the beginning, every sign is changed to its 8bit American Standard Code for Information Interchange (ASCII) format. The outcomes are in the form of 0s and 1s in a string of dimension 12×8=96 of the form:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

(space) I n f o

00100000 01101001 01101110 01100110 01101111

r m a t i o

01110010 01101101 01100001 01110100 01101001 01101111

n

01101110

In the next step, a unique barcode of nineteen bits is sent to the string of 0s and 1s for justification. Let the assumed barcode is 1100110111000110111.

This generates a binary string of length 19+96=115, barcode {11001101110001101110010000001101001011011100110 01100110111101110010011011010110000010111010001101001010 10111101101110}.

Now, every 0 bit is changed into A or C and each1 into T or G. The scheme also balances the GC content and controls the secondary block.

TTACGTATACTAACACAGTCGCCTCTGATTGCAGTCATTC CGGATTGTCTGTAAGCATGCTGCGCGGAACATCTTGAGCC ATGAGCATCGGATGGTCTGATTGC

Lastly,2short strands of RNA or DNA (generally about 18-22 bases) that serve as a starting point for DNA synthesis. of dimension 22 are included in both sides of the DNA structure. The first primer is AGCACATCATAGAGGAATCGAG and latter is CTCGATTCCTCTATGATGTGCT. Therefore, the codeword of DNA is of length 159and is:

AGCACATCATAGAGGAATCGAGTTACGTATACTAA CACAGTCGCCTCTGATTGCAGTCATTCCGGATTGTC TGTAAGCATGCTGCGCGGAACATCTTGAGCCATGA GCATCGGATGGTCTGATTGCCTCGATTCCTCTATGA TGTGCT

ALGORITHM

$X = \text{ENCODE}_{\mathbf{a},\ell}(x)$
begin
1 if $(\ell \geq n)$
2 $t \leftarrow 1$;
3 $y \leftarrow x$;
4 while $(y \geq |\bar{A}_t| G_{n,\ell-t})$
5 $y \leftarrow y - |\bar{A}_t| G_{n,\ell-t}$;
6 $t \leftarrow t + 1$;
7 end;
8 $a \leftarrow \lfloor y/G_{n,\ell-t} \rfloor$;
9 $b \leftarrow y \bmod G_{n,\ell-t}$;
10 return $\mathbf{a}^{(t-1)}\bar{a}_{t,a+1}\text{ENCODE}_{\mathbf{a},\ell-t}(b)$;
11 else
12 return $\theta_\ell(y)$;
13 end;
end;

$x = \text{DECODE}_{\mathbf{a}}(X)$
begin
1 $\ell = \text{length}(X)$;
2 $X = X_1 X_2 \ldots X_\ell$;
3 if $(\ell < n)$
4 return $\theta^{-1}(X)$;
5 else
6 find(s,t) such that $\mathbf{a}^{(t-1)}\bar{a}_{t,s} = X_1 \ldots X_t$;
7 return $\left(\sum_{i=1}^{t-1} |\bar{A}_i| G_{n,\ell-i}\right) + (s-1)G_{n,\ell-t} + \text{DECODE}_{\mathbf{a}}(X_{t+1} \ldots X_\ell)$;
8 end;
end;

5. GOLDMAN et al METHOD

The first step with which Goldman [2] et.al started the work was to convert the digital information into binary representation via ASCII encoding. Then the conversion of each byte into 5 or 6 bits was carried out. Each bit was used to select one out of three oligonucleotides which was different from the last encoded oligonucleotide. This was done so that no homopolymer has length greater than one. The final DNA string was divided into a number of segments. Each segment had a length of 100 oligonucleotides and exhibited the property of overlapping in 75 bases with each adjoining segment. [8]

This is quite beneficial as it covers each base 4 times. The alternate segments were reverse complemented. Goldman et. al encoded a 739 kB sized file containing 154 sonnets of Shakespeare, an excerpt from speech of Martin Luther King 'I have a dream 'in MP3 format, a scientific paper in form of a PDF and a colored photograph of the European Bioinformatics Institute in JPEG 2000 format along with an approximated Shannon information of 5.2 X 106 bits into DNA.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

The number of copies created for each sequence was 1.2 X 107. The error rate was found out to be 1 per 500 bases. The Illumina HiSeq 2000 system was used for sequencing and it was successfully decoded.

An accuracy of 100% was noted when the original data was decoded after a number of postprocessing steps.

Example:

Let the text selected for encoding is “Birney and Goldman”.

First, to compress the data, Huffman coding base 3 is applied, which results in

B i r n e y (space)

S1 = 20100 20210 10101 00021 20001 222111 02212

a n d (space) G o l

01112 00021 22100 02212 222212 02110 02101

d m a n

22100 11021 01112 00021.

• Let n = len (S1) = 92, which is equal to10102 in base 3. Hence, S2 = 00000000000000010102 (an 20 length encoding) and S3 = 0000000000000 (an encoding of length 13). Therefore,

S4=S1S3S2=20100202101010100021200012221110221201112000212210002212222212021002101221001102101112000210000000000000000000000000000001010

of total length 92 + 13 + 20 = 125.

• Differential coding is applied to S5 according to the table

NEXT

PREVIOUS	0	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

[6]

results in an encoding of S4 that reads as

S5=TAGTATATCGACTAGTACAGCGTAGCATCTCGCAGCGAGATACGCTGCTACGCAGCATGCTGTGAGTATCGATGACGAGTGACTCTGTACAGTACGTACGTACGTACGTACGTACGACTAT.

• Since len (S5) = 125, it implies that two DNA blocks are there namely F0 and F1 of length 100 which overlap in exactly 75 bps, i.e.,

F0=TAGTATATCGACTAGTACAGCGTAGCATCTCGCAGCGAGATACGCTGCTACGCAGCATGCTGTGAGTATCGATGACGAGGACTCTGTACAGTACGTACG

and

F1=CATCTCGCAGCGAGATACGCTGCTACGCAGCATGCTGTGAGTATCGATGACGAGTGACTCTGTACAGTACGTACGTACGTACGTAC GTACGTACGACTAT.

Moreover, the odd-numbered DNA blocks are reverse complemented so that

F1=ATAGTCGTACGTACGTACGTACGTACGTACGTACTGTACAGAGTCACTCGTCATCGATACTCACAGCATGCTGCGTAGCAGCGTAT CTCGCTGCGAGATG.

• The file identification for the text equals 12. This gives ID0 = ID1 = 12. The 12 bits intra-file location for F0 equals intra0 = 000000000000 and for F1, it equals intra1 = 000000000001. The parity check P(i) for block F(i) is the sum of the bits at odd locations in ID iintrai taken mod 3. Thus, P0 = P1 = 1+0+0+0+0+0+0 mod 3 ≡ 1. By appending IXi = IDiintraiPi to Fi we get,

F_0’=TAGTATATCGACTAGTACAGCGTAGCATCTCGCAGCGAGATACGCTGCTACGCAGCATGCTGTGAGTATCGATGACGAGTGACTCTGTACAGTACGTACGATACGTACGTACGTC

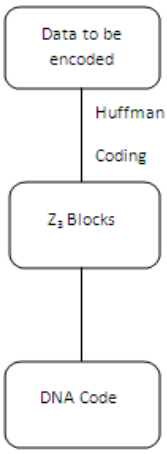
F_1’=ATAGTCGTACGTACGTACGTACGTACGTACGTACTGTACAGAGTCACTCGTCATCGATACTCACAGCATGCTGCGTAGCAGCGTATCTCGCTGCGAGATGATACGTACGTACGAG.

• The last step is to prepend C or G or A or T to odd and even blocks, respectively. The final DNA code words equals

F_0’’=ATAGTATATCGACTAGTACAGCGTAGCATCTCGCAGCGAGATACGCTGCTACGCAGCATGCTGTGAGTATCGATGACGAGTGACTCTGTACAGTACGTACGATACGTACGTACGTCG

F_1’’=TATAGTCGTACGTACGTACGTACGTACGTACGTACTGTACAGAGTCACTCGTCATCGATACTCACAGCATGCTGCGTAGCAGCGTATCTCGCTGCGAGATGATACGTACGTACGAGC.5.1 FLOWCHART

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018



This flowchart [4] depicts the working of the Goldman technique. Data to be encoded is converted into Z_3 blocks using Huffman Coding. The Z_3 blocks are converted to the required Digital code.

Binary digits containing the ASCII codes were converted to base-3 Huff man code that replaced each byte with base-five or six base-3 digits. Each of bit was encoded with one of the three nucleotides different from the previous one used to avoid homo-polymers that caused mistakes in synthesis of DNA. DNA structure was divided into chunks each of length 117 base pair (bp). Huffman code is a specific kind of ideal prefix code that is regularly utilized for lossless information pressure. The way toward finding as well as utilizing such a code continues by methods for Huffman coding, a calculation created by David A. Huffman while he was a Sc.D. understudy at MIT. The yield from Huffman's calculation can be seen as a variable-length code table for encoding a source image, (for example, a character in a document). The calculation gets this table from the evaluated likelihood or recurrence of event (weight) for every conceivable estimation of the source image.

6. CONCLUSION

Unlike most of the other storage devices, DNA does not decompose over a period of time. Just a few grams of DNA can store the world's yearly produced information. The techniques discussed in the paper are the most efficient and practically viable techniques available in the present scenario. Storing data in DNA is profitable as, it is much smaller than conventional media. Also, DNA is not damaged for more than 100 years, which is orders of magnitude more than the conventional media. Given the limitations of Silicon based storage mediums, it is the perfect time to start using bio-chemicals as a storage media. These techniques provide us with a medium to store digital data in DNA. It is the perfect opportunity to for the computer industry to acquire from biotechnology industry for an advancement in computer frameworks.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Dr. Prateek Jain, Accendere Knowledge Management Services., Ms. Renuka Solanki, Ms. Shaveta Malik, FET, MRIIRS for their valuable comments that led to substantial improvements on an earlier version of this manuscript.

REFERENCES

[1] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," Science, vol. 337, pg. 1624,2012.
[2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, information storage in synthesized DNA," Nature, 2013.
[3] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA" Angewandte Chemie International Edition, vol. 54pp. 2551–2556, 2015.
[4] I. S. Reed and G. Solomon, "Polynomial codes", Journal of the society for industrial and applied mathematics, vol. 8, pp. 300–305, 1960.
[5] S. Kosuri and G. M. Church, "Large-scale de novo DNA synthesis" Nature, vol. 11, pp. 498–508, 2014.
[6] S. Roy and M. Caruthers, "Synthesis of DNA and their analogs via phosphoramidite and h-phosphonate chemistries," Molecules, vol. 18, pp. 14 268–14 270, 2013.
[7] J. Tian, H. Gong, N. Sheng, X. Zhou, E. Gulari, X. Gao, and G. Church, "Accurate multiplex gene synthesis from programmable DNA microchips," Nature, vol. 432, no. 7020, pp. 1052–1056, 2004.
[8] S. Kosuri, N. Eroshenko, E. M. LeProust, M. Super, J. Way, J. B. Li, and G. M. Church, "Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips," Nature biotechnology, vol. 28, pp. 1294–1300, 2010.