International Journal of Civil Engineering and Technology (IJCIET)
Volume 8, Issue 12, December 2017, pp. 108–114, Article ID: IJCIET_08_12_013
Available online at http://http://iaeme.com/Home/issue/IJCIET?Volume=8&Issue=12
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

A REVIEW ON VARIOUS ENCODING SCHEMES USED IN DIGITAL DNA DATA STORAGE

A. Swati, Forum Mathuria, S. Bhavani, E. Malathy, R. Mahadevan
School of Information Technology and Engineering
VIT, Vellore, Tamil Nadu, India

ABSTRACT

The limit of existing storage media is diminishing. Interest for information stockpiling is developing exponentially. Ordinary part of information is delivered and this requires high thickness stockpiling gadgets which can hold esteems for quite a while. DNA to chronicle information is an appealing plausibility since it is to a great degree thick. Yet, with regards to taking care of huge information, the information of an organization or of the world in general, the present information stockpiling innovation comes no place close to have the capacity to oversee it proficiently. DNA has been recognized as a potential medium for mystery composing, which accomplishes the route towards DNA cryptography and stenography. DNA used as a natural memory gadget alongside huge information stockpiling and investigation in DNA has prepared towards DNA registering for taking care of computational issues. In this paper, we analyse the different encoding techniques of DNA and also the advantages and disadvantages of the DNA archival system.

Key words: DNA, nucleotides, digital data, encoding.

Cite this Article: A. Swati, Forum Mathuria, S. Bhavani, E.Malathy, R. Mahadevan, A Review on Various Encoding Schemes Used in Digital DNA Data Storage. *International Journal of Civil Engineering and Technology*, 8(12), 2017, pp. 108–114. http://iaeme.com/Home/issue/IJCIET?Volume=8&Issue=12

1. INTRODUCTION

The interest for putting away an ever-increasing number of information is expanding step by step. Alarmingly, the exponential development rate effortlessly surpasses our capacity to store it, notwithstanding when representing gauge upgrades away advances. A huge portion of this information is in documented shape [1]. Since it exceedingly dense data can be filed for quite a while. Information storage and recovery is unavoidable and its conservation issue is approaching over our data arrange. The adventure of information stockpiling started from Rocks, bones, Paper, Punched cards, Magnetic Tapes, Drums, films, Gramophone records, floppies and so forth. Information stockpiling has in the present situation reached out to optical plates including CDs, DVDs, Blu-beam Disks to Portable hard drives and USB streak drives. With the work of advanced frameworks with the end goal of age, transmission, and capacity of data, there rises a requirement for dynamic and progressing upkeep of computerized media. With the monstrous measures of computerized information that must be

put away for some time later, an issue emerges in the capacity of powerful measures of information. As the information expands, the present information stockpiling innovation would not be sufficient to store information in future as information is developing each day. The capacity frameworks as of now utilized are attractive and are optical capacity mediums. Magnetic tape is a medium for attractive account, made of a thin, magnetisable covering on a long, limit piece of plastic film. Magnetic tape altered communicate and recording. It permitted radio, which had dependably been communicated live, to be recorded for later or rehashed airing. Tape remains a practical other option to plate in a few circumstances because of its lower cost per bit. This is an expansive preferred standpoint when managing a lot of information. Analyst's dedication has been driven towards improvement of a capacity instrument which overcomes the previously mentioned disadvantages effectively. Considering the way fossil bones save hereditary material for a long time, scientists gave careful consideration towards utilizing de-oxyribonucleic corrosive (DNA) as a capacity medium.

DNA has a mind-blowing storage limit. Engineered DNA groupings have for some time been viewed as a potential medium for advanced information stockpiling. DNA-based capacity additionally has the advantage of endless significance: insofar as there is DNA-based life, there will be solid motivations to peruse and control DNA. A standout among the most critical favourable circumstances of utilizing DNA as a capacity medium is that the capacity thickness is high. DNA comprises of adenine, guanine, cytosine, and thymine in sets of A-T and C-G. As the dire requirement for high limit stockpiling medium ascents, DNA is viewed as perfect in such manner as single nucleotide can speak to 2 bits of data. There are numerous approaches to reinforcement the information. One can utilize cloud administrations to store information. In any case, to get to information which is put away in a remote cloud, a web association is required constantly. DNA can withstand a more extensive scope of temperatures (−800∘C– 800∘C). It uses control utilization million times more adequately than an advanced PC. Moreover, it benefits more stockpiling choices as it stores information in a nonlinear structure not at all like the vast majority of the media putting away information in a straight structure. As DNA can retain information for centuries, DNA can be used for long-term storage [2].

Due to high density, the DNA can store a large amount of data in very small space [3]. Moreover, this examination presented the possibility that DNA Storage is significantly more private and secure than Digital Storage on Silicon gadgets that too without an express encryption component. The fundamental developments in this examination were the utilization of a mistake rectifying encoding plan to guarantee the to a great degree low information misfortune rate, and in addition encoding the information in a progression of covering short oligonucleotides identifiable through a succession based ordering plan. Likewise, the arrangements of the individual strands of DNA covered such that every area of information was rehashed four times to stay away from blunders. Two of these four strands were developed in reverse, additionally with the objective of killing blunders. Advancement and improvement of encoding models in the current past have been compressed right off the bat in this subsection. Furthermore, encryption plans utilized for encryption of information in DNA have been examined in detail, diagnosing their critical highlights, points of interest, downsides related. Thirdly, a few methodologies utilized for outlining of codons have been talked about. Fundamental information capacity styles have been investigated in a basic way by distinguishing the points of interest over the other.

2. RELATED WORKS

The possibility of Digital Storage in DNA was first in a roundabout way executed in 1999 by Clell and, Risca, Bancroft. They prevailing with regards to putting away encoded words in short DNA strands (Microdots). The innovation was utilized as a part of World War II to impart mystery information. A microdot was a downscaled photo of a wrote page encoded in a period (.) in a safe letter [4].Among early cases of DNA information stockpiling, in 2007 a gadget was made at the University of Arizona [5], utilizing tending to particles to encode bungle destinations inside a DNA strand. These criss-crosses were then ready to be perused out by playing out a limitation process, in this manner recouping the information. This framework has various points of interest over different strategies. Right off the bat, not at all like different techniques in which bespoke atoms are combined for each new DNA encoding, a typical arrangement of particles could be utilized to encode any discretionary information [6].

DNA union is right now costly, and arduous, so this implies this speculation can be utilized to encode various arrangements of information, utilizing a similar arrangement of DNA particles. The encoded DNA made here is likewise "bio-good", implying that, on a fundamental level it can be promptly embedded into, and engendered inside, a life form. The information is changed over to ASCII configuration and after that encoded utilizing Huffman code given what's more, changed over to base-3 organize. The record is figured utilizing length, document ID and equality. This data is changed over to nucleotide organize utilizing characterized table and dodges nucleotide rehashes [7]. Each letter on the console is mapped to a blend of 4 nucleotides. This strategy permits putting away information in half space than a regular advanced stockpiling framework. Indeed, despite the fact that this strategy is very enhanced, it just considers letters on the console. Putting away messages in DNA was first exhibited in 1988 and the biggest venture to date encoded 7920 bits [8].

The little size of past work originates from the trouble of composing and perusing long immaculate DNA groupings, and has constrained more extensive applications (table S1). Here, we build up a procedure to encode subjective computerized data utilizing a novel encoding plan that uses cutting edge DNA union and sequencing innovations. A draft of html coded book with 53,426 words, 11 JPG pictures, and 1 JavaScript program is being encoded. This includes a 5.29 MB stream. Essential focal points of this approach over the earlier approaches utilized for encoding incorporate the work of one-piece portrayal per base (G or T for 1 and An or C for zero). This acquires the capacity of encoding groupings which are trying to be perused or composed because of containing rehashes, optional structures, and outrageous GC content. Difficulties related with this plan incorporate the cost which is unfeasible and the ideal opportunity for perusing and composing onto DNA. However, the cost related with integrating and sequencing of DNA has been dropping at 5– 12 exponential rates every year which is moderately much expedient. These can be utilized later on as DNA sequencers which are hand-held are as of late accessible which attaches the perusing procedure.

3. ENCODING AND DECODING DIGITAL FILES

The general encoding is carried out by binarization of digital file to obtain the binary codes. The obtained binary codes are encoded to nucleaobase codes. The process of Synthesis is carried out to make the DNA sequences. Thus encoding is done on the digital files. To decode the process of sequencing the DNA is followed by decoding. From the DNA sequences, the nucleobase is generated. Later they are converted to binary codes by decoding them using some decoding schemes. Finally the binary codes are read and we obtain the digital files as

the output. We incorporate several encoding and decoding schemes for this conversion. The figure 1 represents the general encoding and decoding procedures in detail.

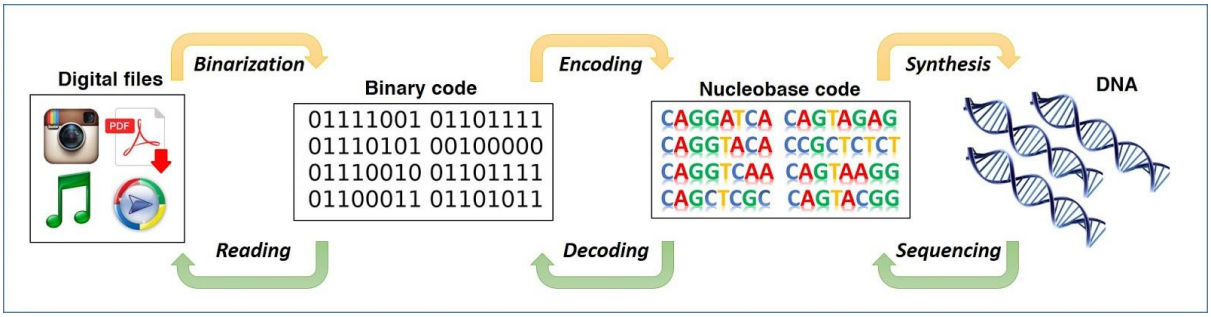

Figure 1 DNA Encoding/ Decoding Procedure [15].

4. TYPES OF ENCODING SCHEMES

Codes for the most part thought about that as an alphabetic dialect is being encoded in DNA. Albeit the majority of the investigates considered English as the alphabetic dialect, it could have been utilized for even shorthand, which is the written work conspire for phonetics. For a code to be ideal it ought to fulfil the double criteria as takes after:

- It should utilize DNA (nucleotides) financially, for the most part since combining of expanded oligonucleotides is a costly procedure however reproducing has all the earmarks of being relatively conservative.
- It ought to have the capacity to recreate the message after encoding of information. One of the essentials for a decent DNA coding strategy is the reasonable utilization of nucleotide bases per character.
- It has been numerically demonstrated that the base to character proportion of around three is most ideal and sparing for a coding framework. This is the motivation behind why numerous specialists favoured (and still keep on doing) the Huffman Coding Scheme.
- As the coding scheme to encode messages into the base 4 DNA form, it had to be modified.

The types of encoding schemes are mentioned in the following.

4.1. The Perfect Genetic Code

In this approach Doig calls attention to the way that through changing the codon length effectiveness of the code can be expanded altogether [9]. Here Doig employments a variable codon length through utilizing more successive amino acids with shorter codon length while uncommon acids are spoken to utilizing a more drawn out codon length. As there is almost no repetition, any change would cause an adjustment in amino acids. Furthermore, if the change is boss to variance of codon length huge numbers of the changes will be broad frame shift transformations. As indicated by Doig, it is incomprehensible for the move towards utilization of flawless hereditary code, considering a basic case that Val has been coded for four codons involving 3 bases. As the third does not pass on any data it is powerful to code Val utilizing two bases as it were. The trouble of hardware to move from settled codon length to variable codon length in addition to the priory specified downsides prompts not utilizing this successful code despite the fact that it expands the productivity through utilizing variable length codons.

4.2. The Alternating Code

This plan comprises of 6 base codons which are 64 in number including pyrimidines and purines. Development of the message DNA in a completely manufactured nature is the essential component of this approach. As this makes completely fake DNA it is appropriate for long haul stockpiling which beats the downside of Huffman code. Likewise, it offers advantages, for example, being isothermal and mistake recognizing be that as it may, it isn't better than comma code. Substituting code likewise involves tedious highlights which make it non economical [10]. It is the primary disadvantage related with this coding scheme. Therefore, consideration of the specialists has been driven towards building up a conservative code without redundant highlights [10].

4.3. Comma-Free Code

It is otherwise called prefix free code. This includes settled length base casings without commas to isolate the edges. In this way, it utilizes a programmed outline location component. Sans comma code does not comprise of indistinguishable four base sets which is the main method for blocking from normal DNA groupings [11]. These codons are conceivable to be perused just in one way and bolster blunder recognition components too. In spite of the fact that without comma code is powerful and the blunder adjustment attempts to rectify against little scale misfortune, for example, DNA point transformations.

4.4. Huffman Coding

This code utilizes the guideline of changing the length of images utilized for speaking to a character [12]. Most intermittently showing up character in the content is relegated the most reduced number of images while the minimum intermittently showing up character is allotted the most number of images [13]. Utilizing this guideline prompts creating of an extremely temperate code. This overcomes the disadvantage of the Huffman code, being restricted just to the letters of the letters in order. This depends on a development of a plasmid library with extraordinarily outlined preliminaries inserted alongside the message for quick recovery.

Table 1 Various Encoding Schemes vs Its Features

	Huffman code	**Comma free code**	**Alternate code**	**genetic code**
Long term storage	-	--	Yes	-
Error correcting	-	Yes	Yes	-
Synthetic DNA	-	-	Yes	-
Economical	Yes	-	-	-
Protection	Yes	-	-	-
Base-to-character ratio	2.2	Variable	6	Variable
Isothermal melting	-	-	Yes	-

Record plasmids contain insights about the structure of the data library [14]. A decent encoding plan ought to have efficient utilization of nucleotide per character which is around 3.5 here (base-to character proportion). The table 1 below shows the comparison among the various encoding schemes for digital DNA storage.

5. CONCLUSIONS

Plainly information storage in DNA is not any more bound to science however is being acknowledged at extremely encouraging rates by inquire about groups everywhere throughout the world. Information is scrambled into DNA utilizing assorted codes and this article breaks down and talks about the codes utilized for encoding information. Various methodologies for outlining DNA codons and differing information stockpiling styles have been investigated in

detail distinguishing the advantages and disadvantages of each approach. As DNA can hold information for a large number of years, it is conceivable to store information for quite a while. By utilizing this procedure, information is compacted and the security to the information is given. Parallel perusing of records is additionally conceivable empowering clients to peruse numerous documents in the meantime. This procedure keeps up two duplicates of information. Henceforth if there should arise an occurrence of information harm, its duplicate can be utilized to peruse information. On account of any blunders while encoding the information, the mistake is limited to that specific document and no other record is influenced because of that blunder. Rather than utilizing traditional capacity gadgets which have less ability to store information, DNA-based capacity technique be utilized as a part of removed future to store information secured way and for long time stockpiling and illuminate the issue of constrained space.

REFERENCES

[1] Laddha, R., & Honwadkar, K. Digital Data Storage on DNA. *International Journal of Computer Applications*, *142*(2), 2016.

[2] C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland. Long Term Storage of Information in DNA. Science, 293, 1763 (2001).

[3] George M. Church, Yuan Gao, Sriram Kosuri. Next Generation Digital Information Storage in DNA. Science, 337, 1628 (2012).

[4] C. T. Clelland, et al., "Hiding messages in DNA microdots," Nature, vol. 399, no. 1033, pp. 533–534, 1999.

[5] C. Bancroft, T. Bowler, B. Bloom, and C. T. Clelland, "Longterm storage of information in DNA," Science, vol. 293, no. 5536, pp. 1763-1765, 2001.

[6] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., & Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, *494*(7435), 77-80.

[7] 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).

[8] E. Kac, "Genesis-art of DNA," 1999, http://www.ekac.org/geninfo .html

[9] A. J. Doig, "Improving the efficiency of the genetic code by varying the codon length—the perfect genetic code," Journal of Theoretical Biology, vol. 188, no. 3, pp. 355–360, 1997

[10] N. Yatchie, Y. Ohashi, and M. Tomita, "Stabilizing synthetic data in the DNA of living organisms," Systems and Synthetic Biology, vol. 2, no. 1-2, pp. 19–25, 2008.

[11] M. Ailenberg and O. D. Rotstein, "An improved Huffman coding method for archiving text, images, and music characters in DNA," BioTechniques, vol. 47, no. 3, pp. 747–754, 2009

[12] G. C. Smith, C. C. Fiddes, J. P. Hawkins, and J. P. L. Cox, "Some possible codes for encrypting data in DNA," Biotechnology Letters, vol. 25, no. 14, pp. 1125–1130, 2003.

[13] D. A. Huffman, "A method for the construction of minimum redundancy codes," Proceedings of the IRE, vol. 40, no. 9, pp. 1098–1101, 1952.

[14] M. K. Rogers and K. C. Seigfried-Spellar, "Digital forensics and cyber crime," in Proceedings of the 4th International ICST Conference on Digital Forensics & Cyber Crime (ICDF2C '12), Lafayette, Ind, USA, October 2012.

[15] http://blog.agupieware.com/2016/04/data-and-dna-encoding-digital-files.html.

[16] Swetha Annangi, RTL Design of Efficient Modified Ru n-Length Encoding Architectures Using Verilog HDL, International Journal of Electronics and Communication Engineering and Technology , 8(1), 2017, pp. 52–57

[17] Md. Ajmal Sadiq, T. Naga Raju and Kumar. Keshamoni, Modeling and Simulation of Test Data Compression Using Verilog, International Journal of Electronics and Communication Engineering & Technology, 4 (5), 2013, pp. 143–141.

[18] Bangaru Kalpana, Amrut Anilrao Purohit and R. Venka ta Siva Reddy, Area Optimization of SPI Module Using Verilog HDL, International Journal of Electronics and Communication Engineering & Technology (IJECET), 7 (3), 2016, pp. 38–45.