Mapana - Journal of Sciences
2023, Vol. 22, Special Issue 2, 319-333
ISSN 0975-3303 | https://doi.org/10.12723/mjs.sp2.18

Unlocking the Future: DNA Encryption for Secure and Efficient Massive Data Storage

Shorya Rawal*, Rudraksh Gohil*, Jayapriya J* and Vinay M*

Abstract

DNA has emerged as a promising medium for digital data storage because of its high density, longevity, as well as energy efficiency. However, the security provided by DNA storage systems remains a concern, particularly as the technology is adopted for sensitive data applications. DNA encryption offers a potential solution to this problem by encoding the stored data in a secure and reversible manner. In this paper, a new DNA encryption for storage applications by editing or creating a new DNA sequence to store big data for archival purposes in an encrypted format to provide security, is proposed. It is concluded that DNA encryption is a promising approach for securing digital data in DNA storage systems, and there is requirement for further research to optimize the performance and reliability of this technology.

Keywords: DNA storage, encryption, encode, digital data, sequence, archival

I. Introduction

DNA storage is a method of storing digital information in the form of DNA molecules. DNA is a highly compact and durable storage medium that stores vast amounts of information in a very small space. One of the key advantages of DNA storage is its longevity – DNA remain stable for thousands of years under the right conditions, making it a potential long-term storage medium. DNA encryption

* Department of Computer Science, CHRIST (Deemed to be University), Bangalore, Karnataka, India; Email: shorya.rawal@bca.christuniversity.in rudraksh.gohil@bca.christuniversity.in, jayapriya.j@christuniversity.in vinay.m@christuniversity.in

refers to the process of encoding digital information into DNA molecules in a way that makes it difficult for unauthorized users to access or decode the information. The use of DNA as a storage medium offers a high level of security, as DNA is a highly stable and durable storage medium, and the use of encryption adds an additional layer of security.

DNA encryption refers to the process of encoding digital information into DNA molecules in a way that makes it difficult for unauthorized users to access or decode the information. The use of DNA as a storage medium offers a high level of security, as DNA is a highly stable and durable storage medium, and the use of encryption adds an additional layer of security. It has several properties that make it an attractive candidate for data storage, including its durability, density, and longevity. It also has a storage density that is several orders of magnitude higher than traditional data storage media, such as hard drives and flash memory. Additionally, DNA is extremely durable and has the potential to last for thousands of years if stored correctly.

II. Objectives

- Using DNA to provide a better storage system for massive data.
- To Develop a secure and efficient encryption algorithm that convert plain text data into a DNA code.
- Encryption and ciphering the data in the nucleotide sequence to provide more security.
- Editing or creating DNA in-vitro to implement the encoded sequence.

III. Flow Through

The article is presented as follows: Following the Introduction, Background is given in Section 4. Section 2 gives the idea of the article as Proposed work. The work is discussed in Section number 5th and 6th. Finally, the conclusion is given in Section number 9th.

IV. Background

Deoxyribonucleic acid, commonly known as DNA, is a polymer that carries genetic information in nearly all living organisms. The

discovery of DNA structure and function has revolutionised the field of genetics, biology, and medicine. DNA is a double-stranded molecule composed of nucleotides. A nucleotide is made up of a molecule of sugar, along with phosphate group and a nitrogenous base. The four types of nitrogenous bases are Adenine, Thymine, Guanine, Cytosine. Also referred to as A, T, G and C. The nitrogenous bases pair up in a predefined manner: A with T and G with C. The sugar and phosphate molecules form the backbone of the DNA polymer, while the nitrogenous bases form the rungs of the ladder. The two strands of DNA are antiparallel, meaning that they run in opposite directions.

DNA storage has become trend of research in past few years because of the increasing amount of digital data being generated and the limitations of current storage technologies in terms of capacity, longevity, and energy efficiency. Traditional storage media such as hard drives, flash drives, and magnetic tapes have limited lifespan and require regular maintenance to prevent data loss. Furthermore, as data continues to grow at an exponential rate, the cost and energy consumption of maintaining data centres and servers becomes a significant challenge for many organisations. It has been estimated that all of the world's data can be stored in one gram of DNA, which makes it a very attractive option for long-term data storage. DNA storage also has the potential to be energy-efficient, as the data is stored at room temperature and requires very little power for maintenance. Another potential application for DNA storage is in areas where traditional storage methods are impractical or impossible. For example, DNA storage could be used to store data in harsh environments such as outer space or in extreme conditions on Earth. DNA storage could also be used to store data for future generations, preserving important information or cultural artefacts for centuries or even millennia.

DNA storage is a rapidly developing field that uses the DNA as a storage mechanism to store digital information. The idea of using DNA as a mechanism of storing data is based on the high-density storage capacity of DNA, which is capable of storing an immense amount of information in a very small space. In recent years, researchers have made significant progress in developing DNA

storage systems, which have the potential to revolutionise data storage.

Number of studies were conducted to explore the usage of DNA as a storage medium. According to 2012 study, researchers from the European Bioinformatics Institute (EBI) demonstrated that it is possible to store digital information in DNA. The researchers encoded a book into DNA and were able to retrieve the information with 100% accuracy.

In another study, researchers from Harvard University encoded a 53,000-word book into DNA, which they were able to retrieve with 99.99% accuracy. The researchers also demonstrated that DNA used to store data for a long time, with the stored information remaining intact for more than 2,000 years. Several companies have also entered the DNA storage market, including Catalog Technologies, which is developing a DNA-based data storage system that store vast amounts of information in a very small space. The company claims that its technology store 1 exabyte of data in a single gram of DNA. Despite the potential of DNA storage, there are still several challenges that need to be addressed. One of the biggest challenges is the cost of DNA synthesis and sequencing, which is still relatively high. In addition, there are concerns about the long-term stability of DNA, which affects by environmental factors such as temperature and humidity.

A. Escherichia coli (E. Coli)

Escherichia coli (E. coli) is a species of bacterium that is frequently found in both human and animal intestines. The circular, double-stranded E. coli's DNA is around 4.6 million base pairs long. Over 4,000 genes are encoded by the DNA, and these genes make the different proteins necessary for the bacteria to grow, reproduce, and perform other tasks. Many studies have been conducted on the DNA of E. coli, and it is frequently employed in genetic studies. As E. coli is simple to grow and control in the lab, it is frequently employed as a model organism. Moreover, researchers have created several ways and instruments for modifying the DNA of E. coli, such as methods for adding or removing genes, introducing mutations, and managing gene expression.

Overall, E. coli DNA is essential to the life of the bacteria, and research on this DNA has improved our understanding of the mechanisms governing protein synthesis, genetic control, and other essential biological processes.

B. CRISPR

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) (2) is a revolutionary gene-editing technology that allows for precise and efficient modification of DNA sequences in living cells. The CRISPR system is based on a natural defence mechanism found in bacteria, which allows them to detect and destroy invading viruses by cutting their DNA.

The two main components of the CRISPR system are the Cas protein and the guide RNA. The Cas protein is responsible as a molecular scissor, cutting the DNA at certain locations determined by the guide RNA. The guide RNA is a short piece of RNA that binds to a targeted DNA sequence, directing the Cas protein to make a cut at the targeted location.

Scientists have adapted the CRISPR system for use in a wide spectrum of applications which includes gene therapy, drug discovery along with agricultural biotechnology. With CRISPR, researchers now easily edit genes with unprecedented precision, allowing them to study the function of specific genes, develop modern treatments for genetic diseases, and create organisms including the required traits.

V. Encryption and Encoding Model

To provide more security for the data stored in DNA, this new model of encryption and encoding is implemented. This model uses base64 to encrypt the source ASCII value, converting it to Base4 number system. After achieving the Base4 equivalent for the encrypted value, a rotational-3 encoding is applied to further encrypt it in the DNA model. For invalid codon sequences, further rot5 encoding is applied.

A. Base-64 Encryption

Using a simple Base 64 encryption model, we can encrypt the source:

Example Text	Example
Encrypted Text	RXhhbXBsZQ==
ASCII	82 88 104 104 98 88 66 115 90 81 61 61

B. Base-4 Encoding

Example Text	Example
Encrypted Text	RXhhbXBsZQ==
ASCII	82 88 104 104 98 88 66 115 90 81 61 61
Base - 4	1102 1120 1220 1220 1202 1120 1002 1303 1122 1101 331 331

Codon encoded using the table given below:

Codon	Numeric Digit
A	0
T	1
G	2
C	3

Table 1: Encoding Basic Genetic Code into Base 4 System

Codon Encoded: TTAG TGGA TGAG TTGA TAAG TCAC TTGG TTAT CCT CCT

C. Rotational-3 encoding

The ROT3 or the rotational-3 encoding model uses a simple method to rotate the codon by 3, which is shown below:

Codon	**Rotated Codon**
A	C
T	A
G	T

C	G

Table 2: Encoding Basic Genetic Code into Rot3 System So, using this table, we can encode it even further.

ROT3: AACT AATC ATTC ATCT AATC ACCT AGCG AATT AACA GGA GGA

Now, divide the codon string into parts of 3 to define the biological names for those specific codon sequences.

Encoded Sequence	AACTAATCATTCATTCATCTAATCACCT AGCGAATTAACAGGAGGA
Converts to	ACC TAA TCA TTC ATT CAT CTA ATC ACC TAG CGA ATT AAC AGG AGG

Now, using the table given below, give the codon their specific names.

1st letter	Second Letter: U		Second Letter: C		Second Letter: A		Second Letter: G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Codon	Name	Codon	Name
AAC	Asn	ACC	Stop
TAA	Stop	TAG	Arg
TCA	Ser	CGA	Iie
TTC	Phe	ATT	Asn
ATT	Iie	AAC	Arg
CAT	His	AGG	Arg
CTA	Leu	AGG	
ATC	Iie		

Table 3: Codon Table

VI. Invalid Codon Exception Handling

The invalid codes, UAA, UAG, UGA and AUG, which codes for the stop and the start codons cannot be used for storing data, so the data which is coding for this will be put through a ROT5 encoding model, which is given below:

Codon	Corrected
A	T
T	G
G	C
C	A

Table 4: Codon Correction Table

(UAA) TAA - GTT (GUU)
(UAG) TAG - GTC (GUC)
(AUG) ATG - TGC (UGC)
(TGA) UGA - GCT (GCU)

Table 5: Corrected Codon Sequence Table

VII. Storage

The proceeding crucial step is storing the data in the physical DNA to store and archive the data for a much longer time frame, after finding the encoded data pattern in a codon manner, it is implanted into a DNA molecule. Using the CRISPR technology (3), the base pairs is be implanted.

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) is a revolutionary gene-editing technology that allows for precise as well as efficient modification in sequences of DNA in living cells. The CRISPR system is based on a natural defence mechanism found in bacteria, which allows them to detect and destroy invading viruses by cutting their DNA. The two main components of the CRISPR system are the Cas protein and the guide RNA. The Cas protein acts as a molecular scissor, cutting the DNA at specific locations determined by the guide RNA. The guide RNA is a short piece of RNA that binds to a targeted DNA sequence, directing the Cas protein to make a cut at the targeted location.

Editing the DNA to Store Data

Using the bacteria DNA for now would be enough to demonstrate the current use case of storing the encoded data in the physically DNA. For this demonstration the DNA would be cut and edited using CRISPR to incorporate the needed data encoded in DNA sequences. The steps in CRISPR are:

1) Designing guide RNA (gRNA): A specific gRNA is designed to target the desired DNA sequence for editing. The gRNA contains a sequence that matches the DNA sequence to be cut and a short "spacer" sequence that binds to the Cas9 protein.
2) Assembly of CRISPR/Cas9 complex: The Cas9 protein is a nuclease that cuts DNA, and it binds to the gRNA. Together, the Cas9-gRNA complex can recognize and bind to the target DNA sequence
3) Targeting and cutting the DNA: The Cas9-gRNA complex is introduced into cells, where it scans the DNA for sequences that match the gRNA. When the complex finds a match, it cuts the DNA at that site.

4) DNA repair: After the cutting of DNA there is an attempt by the cells repair mechanism to repair DNA it can be done by 2 methods either nonhomologous end joining (NHEJ) or homology directed repair (HDR). NHEJ often results in small insertions or deletions (indels) at the cut site, which can disrupt the function of the targeted gene. HDR, on the other hand, uses a template DNA sequence to repair the cut, allowing for precise editing of the DNA sequence.
5) Verification: The edited DNA sequence can be verified using various methods such as DNA sequencing or restriction enzyme digestion.

Creating Deoxyribonucleic Acid In-Vitro

The previous part of this research paper talks about editing the nucleotides in the present DNA to store the encoded data, this way of editing genomic sequences can take a lot of trial and error and raise some ethical arguments which stands against genetic information of living organisms. To tackle this argument, the in-vitro synthesis of DNA can be implemented to create a DNA molecule which will only store the given data. The DNA is first made as a simulated model in a DNA Synthesiser.

After creating the needed sequences, the DNA can be synthesised in-vitro using either PCR or Phosphoramidite method.

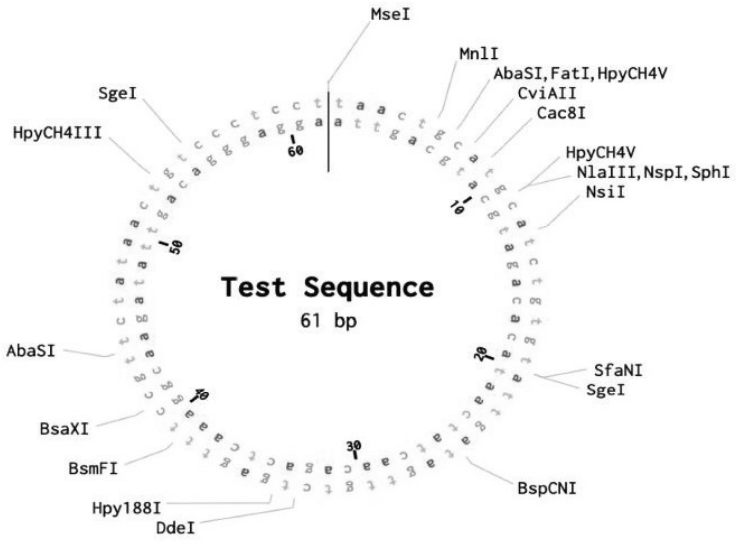

Figure 1: Test Sequence

In-Vitro DNA synthesis using phosphoramidite involves several key steps, including:

- Starting with a solid support, a short strand of DNA called a "primer" is attached to the support. The primer provides a starting point for the synthesis of the new DNA strand.
- Phosphoramidite is added to the reaction mixture. The phosphoramidite is a protected nucleotide that has a reactive group that can bond with the 3' end of the growing DNA chain.
- A chemical reaction is used to remove the protecting group from the phosphoramidite, allowing it to bond with the 3' end of the growing DNA chain.
- Unreacted phosphoramidite is washed away, and any remaining protecting groups are removed from the newly added nucleotide.
- Steps 2-4 are repeated, one nucleotide at a time, until the desired DNA sequence is synthesised.
- The completed DNA strand is cleaved from the solid support and deprotected, resulting in a purified DNA molecule.

In vitro DNA synthesis using PCR (polymerase chain reaction) involves several key steps, including:

1. Denaturation: The double-stranded DNA template is denatured by heating to a high temperature (usually around 95°C), separating the two strands of the template DNA.
2. Annealing: The temperature is lowered to allow the primers to anneal (bind) to the single-stranded DNA template at specific regions on both strands.
3. Extension: A heat-stable DNA polymerase enzyme (such as Taq polymerase) synthesises new DNA strands by extending from the primers along the template DNA, creating complementary strands. The temperature is typically raised to around 72°C for optimal Taq polymerase activity.

4. Repeat cycles: The above three steps are repeated for multiple cycles (usually 20-40 cycles) to exponentially amplify the target DNA sequence.

The exact temperature and time conditions for each step depend on the specific PCR protocol being used and the characteristics of the DNA template and primers.

After getting the synthesized DNA, store it in test tubes or freeze the DNA containing solution to store that data until further use. To view the synthesized DNA, some tools is used to represent the helical structure including the base pairs of the data, the encoded data used in this research paper in simulated and shown below:

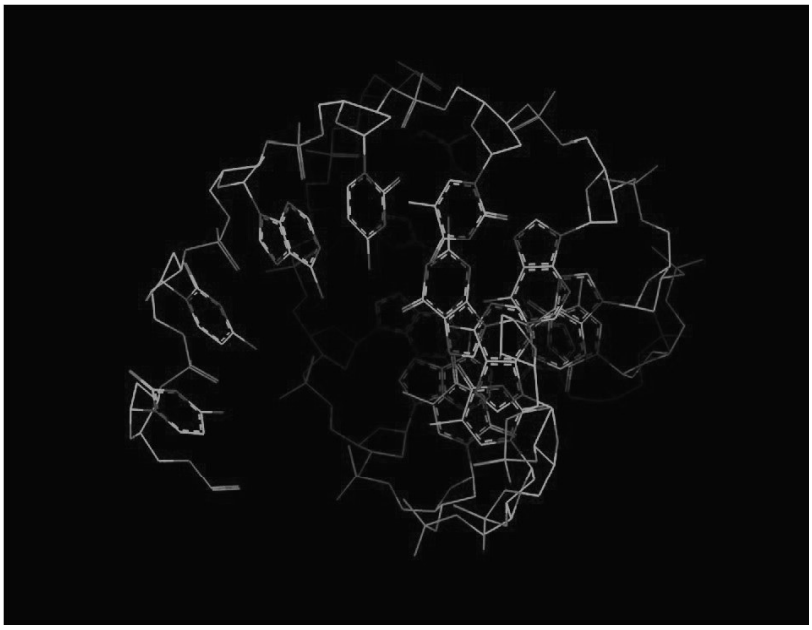

Figure 2: Simulated DNA

VIII. Discussion

The need for a new DNA storage model arises due to the limitations of existing data storage technologies. Traditional storage media such as hard drives, flash drives, and optical discs have limited storage capacity and are susceptible to degradation and data loss over time. In contrast, DNA offers several advantages as a storage medium, including its high density, durability, and longevity [1]. However, current DNA storage methods involve synthesizing and sequencing short DNA fragments, which are time-consuming and costly. In

addition, there is a lack of standardized protocols for DNA storage and retrieval, which hinder the widespread adoption of DNA storage as a viable data storage solution. To tackle this problem this research paper brought out a new method to use Base4 configuration and encoding to securely safeguard data in a DNA molecule along with a encryption model which provides even more security using encryption and cipher methods on the source and the encoded genes, respectively.

Although the usage of DNA as a digital information storage medium for storing digital information has shown a lot of potential, the technology is still in its infancy, and further study is required to evaluate its efficacy as a workable data storage method. Further information on the effectiveness of DNA storage may be found in the following statistics and citations:

- Storage Capacity: DNA storage has a high storage density, which is one of its benefits. In a study that was published in the journal Nature, scientists were able to fit 215 petabytes of data onto one gramme of DNA [4]. As a result, DNA has the capacity to store a significant quantity of info in a comparatively little amount of space.
- Durability: Under the appropriate circumstances, DNA molecules are known to be stable and persist for thousands of years. In a study that was published in Scientific Reports, complete DNA was retrieved from 700,000-year-old horse bones [5]. This shows that DNA may one day serve as a durable medium for storing digital data.
- Cost: Compared to alternative storage options, the cost of producing and analysing DNA is still rather expensive. The National Human Genome Research Institute said that it costs roughly $7,000 to synthesise one mega base of DNA, which is comparable to about one minute of music [6]. In comparison to alternative storage options like hard drives and flash memory, this price is much greater.
- Speed and Efficiency: Reading and writing data to DNA quickly and effectively still poses significant difficulties. It takes a long time to recover data from DNA, and current techniques of encoding data into DNA are sluggish. Research

that appeared in science reported that it required many days to create DNA with a 2.14-megabyte message and then extract the message [7].

Overall, research is still ongoing to determine if DNA storage is a workable alternative for data storage. Although having the capacity to store enormous quantities of data in a tiny area and having a lifespan of thousands of years, DNA storage currently has numerous drawbacks, including expensive prices and sluggish read and write speeds. The speed, effectiveness, and affordability of DNA storage are now being improved via research and development, which might one day make it a workable option for storing massive amounts of data.

IX. Conclusion

DNA encryption storage approach is a promising technology that has the potential to revolutionize data storage. With its high storage density and durability, DNA provide an efficient and long-lasting solution for data archiving. However, the technology is still in its initial stages and there are various challenges that can be overcome, which includes challenges like the high cost of synthesis and sequencing, error rates, and the need for specialized equipment. After the past study this article gives a new approach for storing data in DNA format with encryption model. Furthermore, ethical and privacy concerns related to the use of DNA for data storage need to be carefully considered. Nevertheless, with continued research and development, DNA encryption storage approach has become a viable alternative to traditional storage technologies and may lead to a new era of data storage and management

X. References

[1]. Cao, B., Zhang, X., Cui, S. et al. Adaptive coding for DNA storage with high storage density and low coverage. npj Syst Biol Appl 8, 23 (2022). https://doi.org/10.1038/s41540-022-00233-w J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2]. "CRISPR-Cas systems for editing, regulating and targeting genomes" by Feng Zhang and colleagues, published in Nature Biotechnology in 2014. K. Elissa, “Title of paper if known,” unpublished.

[3]. Digital data storage on DNA tape using CRISPR base editors, Afsaneh Sadremomtaz, Robert F. Glass, Jorge Eduardo Guerrero, Dennis R.LaJeunesse, Eric A. Josephs, Reza Zadegan.

[4]. Church, G. M., et al. (2012). Next-generation digital information storage in DNA. Science, 337(6102), 1628-1628. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[5]. Orlando, L., et al. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature, 499(7456), 74-78.

[6]. National Human Genome Research Institute. (2017). DNA Synthesis Costs. Retrieved from https://www.genome/ about-genomics/factsheets/DNA-Synthesis-Costs.

[7]. Goldman, N., et al. (2013). Towards practical, highcapacity, low-maintenance information storage in synthesized DNA. Nature, 494(7435), 77-80.