A PROPOSED DESIGN AND IMPLEMENTATION OF DE-OXY DRIVE

ARTICLE ON DE-OXY DRIVE
(GROUP PROJECT)

Submitted by

SANCHALI DESHMUKH-20BCE10414
ABHINAY GARG - 20BCE10841
SAMBHAV MEHTA -20BCE11033
ANMOL VERMA - 20BCE11036
SHIVANSH RASTOGI - 20BCE11104

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT BHOPAL UNIVERSITY KOTHRIKALAN, SEHORE

MADHYA PRADESH - 466114

October 2, 2022

ACKNOWLEDGEMENT

We wish to express our heartfelt gratitude to Dr. E. Nirmala, School of Computing Science and Engineering for much of her valuable support and encouragement in carrying out this work.

We would like to thank her for continually guiding and actively participating in our project, giving valuable suggestions to complete the project work, and motivating us to finish the work

ABSTRACT

In the modern world, one of the biggest challenges to solve is data storage. With modernization and technological advancements, data storage comes out as one of the most critical factors to be worried about. Today many big organizations are providing their cloud storage to customers to store their data online but is the space sufficient for any user?
Even if we look at big organizations, they are facing this problem themselves too. The company must store its data over huge servers with a lot of disk drives while consuming a lot of space on the cloud.
Even though cloud storage is an excellent alternative to this problem, there are still many tasks that require the files to be stored in the computer memory itself.

Here comes the role of **Deoxy Drive**. We provide an efficient way of storage to the user where the user can store their data in the form of ***DNA*** and can retrieve the data whenever and wherever required. This is the same as compressing files using the zip files but in a much more effective way. **By storing data in DNA Form, users can store *1 TB* of their data in a nail-sized chip**.
For now, we plan to implement a website where users will be able to upload its file, the site will have the required functionality to store that files in DNA format in a compressed manner. Whenever the user requires this file, they can extract this using their dashboard and the file will take back its original format.

The advantages are as follows:

1. Users will be able to store a huge amount of data in a small place
2. The data which is not of frequent use can be kept easily on this site for any period.
3. The site provides intense security as the files stored in DNA format cannot be read easily without decrypting also for now it is almost impossible to rewrite DNA.

The disadvantages are as follows:

1. Users will be able to upload data, but the uploading process will be done using DNA writing which is a lengthy process
2. Due to a delay in the uploading process, the user will not be able to access data immediately.
3. Currently, there is no advanced technology available to process multiple DNA write queries at the same time
4. DNA storage requires Dark and cold storage
5. This storage medium is in its initial phase; therefore, it will take time to get affordable.

TABLE OF CONTENTS

1. INTRODUCTION

: Introduction

Given how pervasive data storage is, one could assume there are not any significant obstacles for such a crucial component of the majority of enterprises. It is a significant problem to overcome, but you will not succeed unless you start with a solid, long-term data storage solution because you can't analyze without a place to store it. Although it might seem quite simple, the COVID-19 pandemic has made data storage difficult. Some of the most data storage challenges faced in recent years are data security, scalability, data accessibility, and protection.

- Data security

Security breaches have increased in recent years along with the growth in ransomware attacks' frequency and complexity. Attacks like these cause issues for an entire organization. Network perimeter security is the first line of protection for data storage, but there is always a chance that a worker with the right access rights may obtain secure data, use it, and possibly corrupt or destroy it. Storage specialists should make sure that their solution encrypts data in motion and at rest because this possibility, whether intentional or unintentional, should be considered. This will always keep business-critical information secure.

- Scalability

Any data storage solution must have the capability to expand storage space as your business expands. Storage components should be able to scale up or down as necessary to meet changing business needs. Circuit board upgrades for servers, extra servers or standalone storage units, storage through a different data center, or third-party managed storage providers in the cloud, can all be used to achieve this. A cost-effective alternative is third-party storage providers because they can frequently simply supply scalability without the need for expenditures in one rack, floor space, or storage devices.

- Data Accessibility

It is vital that an organization can securely and swiftly restore the data and resources required to resume regular day-to-day business operations in the event of a disaster, be it a ransomware attack or a power outage. Implementing security measures like encryption within the data storage platform is the best method to guarantee that IT personnel will be able to accomplish this after a disaster occurs.

- Data Protection

One of the main purposes of a data storage platform is to let users to ess data whenever they need to without having to fear that it has been accidentally destroyed, changed, stolen, or corrupted. Data protection software applications can guarantee that all data will be accessible in its original form by preventing any potential adjustments to stored

data. Furthermore, if a company maintains data in cold storage, it can employ an archive for possible future retrieval. A variety of apps can also be used to help destroy the data or the storage device if an organization decides it no longer needs a particular piece of data.

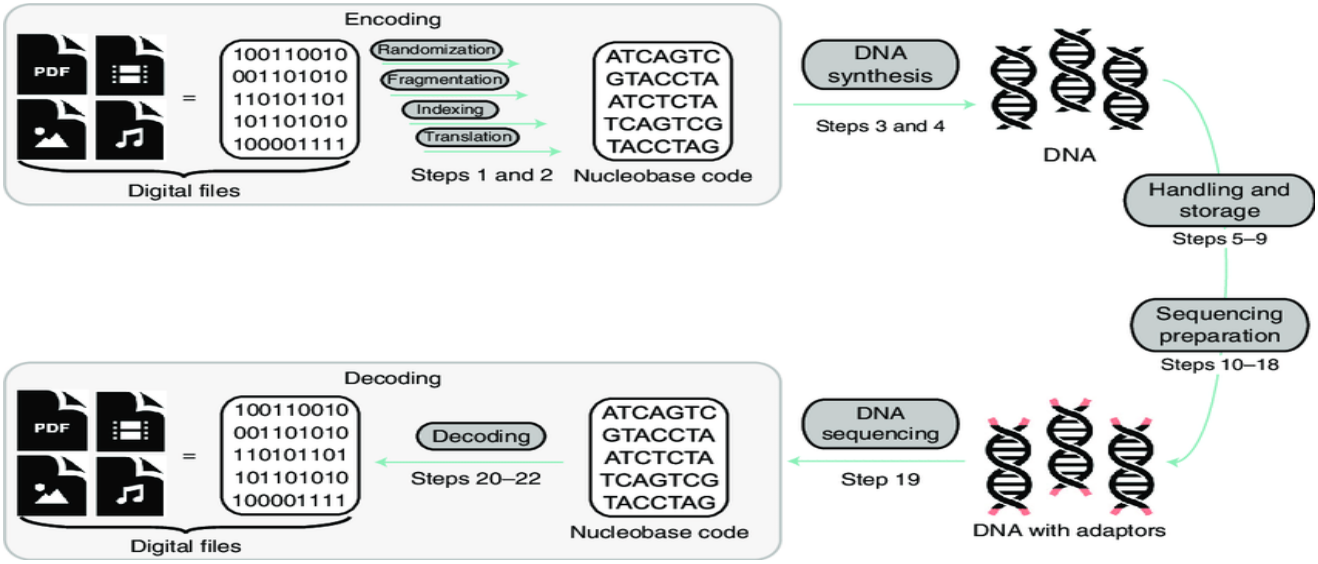
: Motivation for the work

Storage of research data must allow for data exchange, quick access when needed, and easy finding for data management to be done properly. It must be safe from loss and allow others to reuse it. All of this begins with having an effective data management policy and the appropriate tools (storage features), which should be compatible and represent your policy. This is where De-Oxy Drive comes into the picture, it provides users with an effective method of data storage that allows them to save their information in the form of DNA and access it whenever and wherever they need to.

One must make sure that you have a data management policy that incorporates best practices and a compatible data archiving technology if you want to guarantee the long-term preservation of your research data. This provides you with the resources to follow processes and ensures that your data is secure and will be accessible in the future. This motivates us to work on our project which aims at accomplishing the goal of a secure and easy storage system.

: About Introduction to the project including techniques

DNA is a potential storage media that can store and archive the vast amounts of data that exist today.



Process overview of DNA data storage.

In this process, the protocol generally followed includes the steps of encoding digital files by mapping of nucleotides, DNA synthesis which is the writing of the code, then comes the handling and storage of these produced DNAs which in turn is followed by sequencing and DNA decoding, i.e., mapping of required nucleotides back to bits.

During the encoding process, the bits are fragmented using error-correcting codes, then the bits are mapped to nucleotides. The adaptors needed for sequencing are appended to the DNA strands using a two-step technique for preparing the sequencing library. The amplified DNA pool is sequenced, and the original digital file is obtained by decoding it using error-correcting codes.

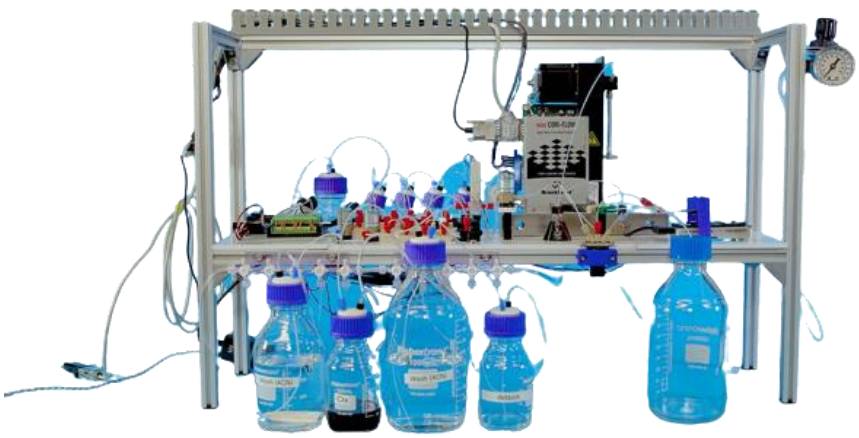
The information provided d by the user must first be converted into a sequence of four bases in the DNA molecules to use DNA molecules for information storage. Any digital information can be easily converted and encoded into the DNA molecule. All sorts of data that can be kept on a hard drive are covered by this.

2. RELATED WORK INVESTIGATION

Existing Work with its Architecture

At present, there is much research going on on this topic by people from different institutes around the world. As any achievement under DNA data storage can bring major changes in the way we store data today people are now taking it seriously.

Currently, countries like China and Russia are showing their interest in it, and many research laboratories are dedicated to DNA data storage. They try to optimize and come up will a better way to store and retrieve data. Some of the research machines look like this:



The above machine is an automated DNA storage device built by Microsoft and the University of Washington
At first, they try to encode the word hello in DNA and then read it out. It may sound like a simple task, but it was the base and showed how the system works. Using this method exponentially increases the density of data storage. Researcher says you can store billions and billions of bytes of data in a cubic inch of DNA.

This software translates digital code into DNA code. Then the code is auto-sent to a synthesizer that combines the required chemicals and liquids, in the correct proportion.
Then we get our DNA molecules into a storage vessel. For reading out the data, the DNA is taken in and we add chemicals, and we push it into the DNA sequencing machine. It automatically converts it into ones and zeros of digital data.

Microsoft and UW also created a programmable system that can move fluid around with a device that is a digital

microfluidic device. The OS used is known as Puddle
It is an open-source OS, made for microfluidic systems and their convenience. Also, Linux is used as an electronic computing system.

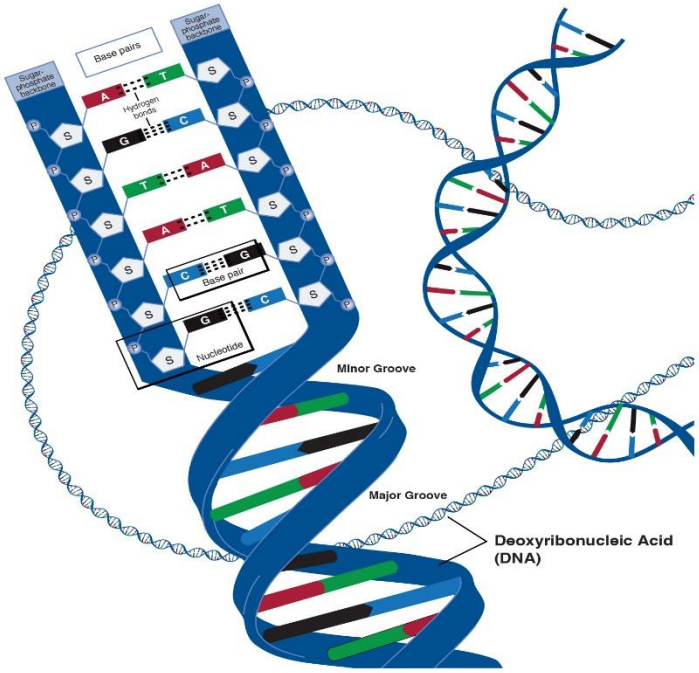
Here's a sample of the Puddle code:

```
a = input(substance_A)

b = input(substance_B)

ab = mix(a, b)

while get_pH(ab) > 7:

 heat(ab)

acidify(ab)
```

This runs in Puddle OS, then the operation is performed on the fluid.

Proposed Work:

DNA Drive: What is it? Thus, DNA, also known as deoxyribonucleic acid, is a potential storage medium that can store and archive the vast amount of data that we produce. Basic data encoding is accomplished by transforming bits into bases.
What does the term encoding mean? The most general definition of encoding is the conversion of data from one form to the next.
All types of data, including photos, video, records, and audio, can be encoded using the Deoxy Drive. which is made possible by converting plain text to a binary representation, which will be further incorporated into the DNA sequence.
It is necessary to convert the text into a binary sequence before it can be mapped to a nucleotide sequence.



Source: https://www.genome.gov/sites/default/files/tg/en/illustration/dna_deoxyribonucleic_acid_adv.jpg

A nucleotide is a substance made composed of a nitrogenous base, a phosphate group, and pentose (5 carbon atoms) sugar.
The four nitrogenous bases in DNA are guanine, cytosine, thymine, and adenine. Their sequence will contain all genetic information and protein-coding. Covalent bonds are formed between the phosphate group of one nucleotide and the third carbon atom of the pentose sugar in the following nucleotide, tying the two together. Thus, sugar-phosphate, sugar—phosphate, and so on make up the DNA backbone created by the pairing of nucleotides. Two coiled polynucleotide chains from this sequence result in the famous double helix of DNA. The entire program is built on a straightforward mapping, in which each nucleotide, represented by a single letter, is represented by two bits of a sequence.

Below given is how the project works:

Step 1: Let us start by taking a simple text and then converting it to the DNA sequence. Here we are taking a variable normal_str to refer it to the normal text entered by the user. The line below will ask for user input.

normal_str = input(“Please input your data.”)

Step 2: The following text entered by the user will be encoded to UTF-8, which represents every character in 1, 2, 3, or 4 bytes. This will help us to convert the given input to a binary sequence using the built-in format method. The binary sequence generated in this case will have all leading 0s with a total of 8 bits. We also use some in-built string method (string. join) to convert the sequence to string format. This is represented by the variable

binary_str = ‘’.join(format(x, 08b) for x in bytearray(original_str, ‘utf-8’))

Step 3: We also know that to convert into nucleotides we need to have a bit length of 2. Hence we convert the binary string to a list where each item has a length of 2.

binary_list = [binary_str[i : i + 2] for i in range(0, len(binary_str), 2)]

Step 4: We already know the 4 nucleotides. Hence we will make a dictionary where every key will represent a 2-bit value of binary_seq and its value as one of the corresponding nucleotides.
An empty DNA_seq is made to which the encoded values are pushed. 2 loops are used for iteration, 1 to iterate in binary_seq and the other to match the corresponding value to the dictionary.
In the below-given code

```
DNA_dict = {
 "00": "A",
 "01": "G",
 "10": "C",
 "11": "T"
}
DNA_list = []
for num in binary_seq
        for key in list(DNA_dict.keys()):
                if num == key:
                        DNA_seq.append(DNA_dict.get(key))
DNA_str = "".join(DNA_seq)
```

So, this was all the code that needed to be used for converting the text. The same can be used for other formats as well. But for any other format of the file, the file 1st needed to be converted to its equivalent text format and then the same to be followed.

Comparative Result Analysis

Due to its amazing high-capacity, high-storage-density properties, and long-lasting potential to store data for thousands of years, deoxyribonucleic acid (DNA) is quickly emerging as a serious medium for long-term archival data storage. To maintain data integrity and retain digital information in DNA, a variety of encoding techniques are often needed.

The comparison of several techniques is as follows:

1. Achieving the Capacity of a DNA Storage with Linear Coding Schemes

The channel output in this model shows a random sample of noisy copies of each sequence. Recent studies have assessed this type of DNA storage channel's capacity under various noise and sequencing models, depending on very advanced typicality-based methods for viability.

In this case, a binary erasure channel-induced noise corruption scenario is used to investigate a multi-draw DNA storage channel. It demonstrates that in this situation, linear coding techniques are used to attain the capacity. In comparison to previous results in the literature, this results in a far simpler derivation of the capacity expression of a multi-draw DNA storage channel.

2. Concatenated Codes for Multiple Reads of a DNA Sequence

This model uses a concatenated coding technique that combines an inner convolutional code with either a time-varying block code or an outside nonbinary low-density parity-check code. For inference from multiple received sequences, two innovative decoding algorithms are proposed; both combine the inner code and channel to a joint hidden Markov model to denote and infer a posteriori probability (APPs).

While the second decoder approximates the APPs by merging the results of individually decoded received sequences, the first decoder computes the precise APPs by jointly decoding the received sequences. The effectiveness of decoding multiple received sequences using the suggested algorithms is assessed using feasible information rates and Monte Carlo simulations.

3. Scaling DNA data storage with nanoscale electrode wells

In this model, because of its density, copy ability, sustainability, and endurance, synthetic DNA is a desirable medium for long-term data storage. New encoding algorithms, automation, preservation, and sequencing technologies have been the focus of recent developments. Despite advancements in these areas, the write throughput, which restricts data storage capacity, continues to be the most difficult barrier to the widespread adoption of DNA data storage.

To overcome the limitations of current DNA synthesis arrays, this model created the first nanoscale DNA storage writer, which anticipates scaling DNA write density to 25 106 sequences per square centimeter. This demonstrates the successful writing and decoding of a message in DNA while limiting DNA synthesis to a region smaller than 1

square millimeter and parallelized over millions of nanoelectrode wells. A realistic DNA data storage system will only be possible with DNA synthesis on this scale, which will enable write throughputs to reach megabytes per second.

4. Coded Shotgun Sequencing

The majority of DNA sequencing methods use this shotgun paradigm methodology, in which numerous short reads are collected from a DNA sequence's unknown places at random.
The high-capacity benefits that could be made if data could be stored in long DNA molecules rather than short-length ones are shown by this model. It advocates taking advantage of overlaps, which are typically present in shotgun-sequenced readings, and enable capacity improvements.

It also shows how we can cut the minimum needed read length and the number of reading samples by a factor of log n while still allowing for arbitrarily close to perfect reconstruction when we shotgun sequence a DNA string that is a codeword from a codebook. the creation and preservation of DNA in general admit errors that can cause bit flips and erasures.

5. The DNA Storage Channel: Capacity and Error Probability

The DNA storage channel is taken into account, in which the M Deoxyribonucleic Acid (DNA) molecules that make up each codeword are first stored randomly, followed by N replacement samples, before being sampled N times and finally sequenced across a discrete memoryless channel. Lower (achievability) and higher (converse) bounds on the channel's capacity as well as a lower (achievability) bound on the reliability function of the channel are given for a constant coverage depth M/N and molecule length scaling Theta (log M).

The lower and upper limitations on the capacity generalize a bound that was previously known to only apply to the binary symmetric sequencing channel, and only under specific constraints on the molecule length scaling and the crossover probability parameters. These limitations are entirely lifted for the lower bound and considerably loosened for the upper bound when specified to binary symmetric sequencing channel.

The lower constraint on the reliability function, which is attained under a universal decoder, demonstrates that outages, or errors in which the capacity of the channel caused by the DNA molecule prevents the sampling operation from supporting the goal rate, are the most common error events.

6. Trellis BMA: Coded Trace Reconstruction on IDS Channels for DNA Storage

When a DNA strand is sequenced as part of the reading process in DNA storage, numerous noisy duplicates are created. By combining these copies, a method known as trace reconstruction can yield improved estimations of the original strand. By adding redundancy in the appropriate order, one can further minimize the error rate. This process is known as coded trace reconstruction. For coded trace reconstruction, it is necessary to create both low-complexity decoding algorithms and an insertion-deletion-substitution (IDS) channel model for the DNA storage channel.

As a result, Trellis BMA, a new reconstruction algorithm whose complexity is linear in the number of traces, is developed for this problem and its performance is compared to that of other algorithms.
The model's outcomes demonstrate that it lowers the mistake rate on both simulated and experimental data.

3. CONCLUSION AND RECOMMENDATION

Conclusion

So, lastly, our group has decided to work on the topic of DNA drives which is a storage device that is capable of encoding any form of data, in which we can store billions and billions of bytes of data in a cubic inch of DNA. The drive works as a storage vessel which solves the biggest problem faced by the world today, i.e., storage of information.

Deoxyribonucleic acid emerged as a high potential for long-term active storage medium because of its high capacity, high storage density characteristics, and lasting ability to store data for thousands and thousands of years. As, for now, the DNA drives are in their initial phase so it would not easily available nor it's affordable for the common folk, the uploading process is very lengthy, the time taken to access the data is also high, and it can't be stored anywhere.

Future enhancement

Since 1990, the cost of DNA synthesis and sequencing has by almost 10 million folds, and by looking at the near future the trend will most likely meet the needs of practical DNA storage. According to the prediction made by the Molecular Information Storage Program, the cost of DNA synthesis will reduce to $\$10^{-10}$/bp by the end of 2023. To compete with the current information read and write speeds DNA synthesis and sequencing technologies need improvement of 7-8 orders of magnitude. Also, techniques for erasing and rewriting information in DNA need to be developed. As the current DNA storage support only one-time storage they are right now suitable for storing information such as government documents and historical archives or any other information that does not need to be modified. But a recent development in this field of synthetic biology has shown the possibility of solving this problem. Still, the scope of research in this domain is quite big and we would like to continue our work on the same.

Final Words

As of the end, we would like to Nirmala Ma'am for giving us this opportunity to work on this project.

CITATION

- Agrawal, Rajeev & Nyamful, Christopher. (2016). Challenges of big data storage and management. Global Journal of Information Technology. 6. 10.18844/gjit.v6i1.383.

- Yiming, Dong & Sun, Fajia & Ping, Zhi & Ouyang, Qi & Qian, Long. (2020). DNA storage: Research landscape and prospects. National Science Review. 7. 1092-1107. 10.1093/nsr/nwaa007.

- Silva, Pavani & Ganegoda, Upeksha. (2016). New Trends of Digital Data Storage in DNA. BioMed Research International. 2016. 1-14. 10.1155/2016/8072463.

- Reinsel, D., Gantz, J. & Rydning, J. Data age 2025: the digitization of the world from edge to core. IDC White Paper Doc US44413318 1–29 (2018).

PLAGIARISM CHECKER