Rewritable Random Access DNA Based Storage System

- Guided by Prof. Manish Kumar Gupta

“DNA is like a program but far, far more advanced than any software ever created.”

—Bill Gates, The Road Ahead

Our Team

Nisarg Nampurkar

201901188

Jemish Variya

201901112

Harikrishna Patel

201901212

201901457 Ashray Kothari

Raj Patel 201901306

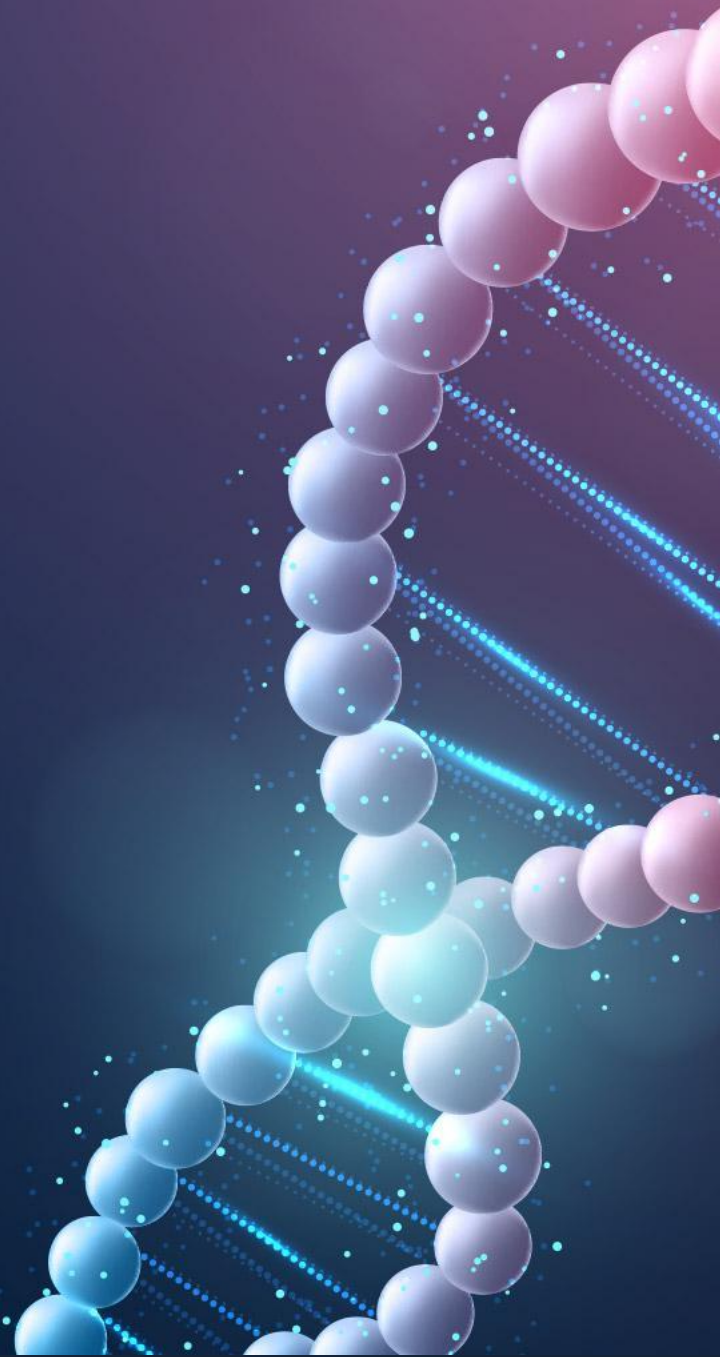
Agenda

1 Introduction

2 Literature Review

3 About Our System

4 References

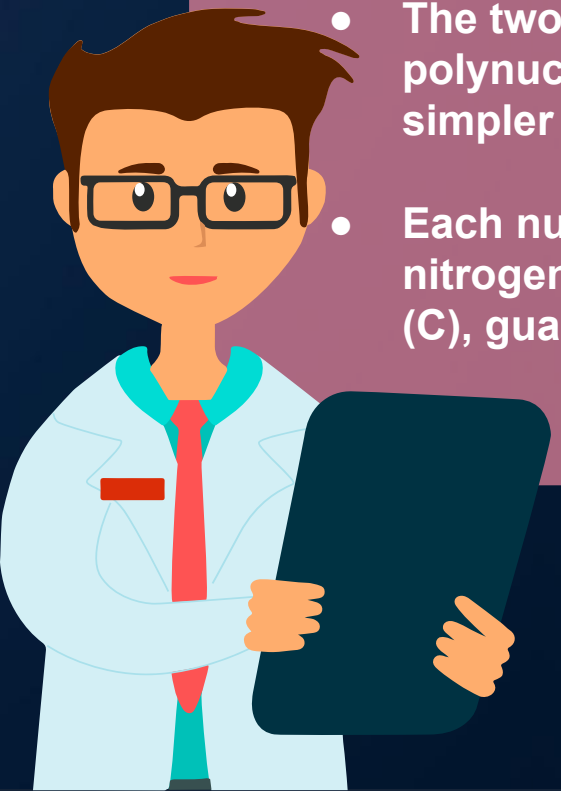


1

Introduction

DNA

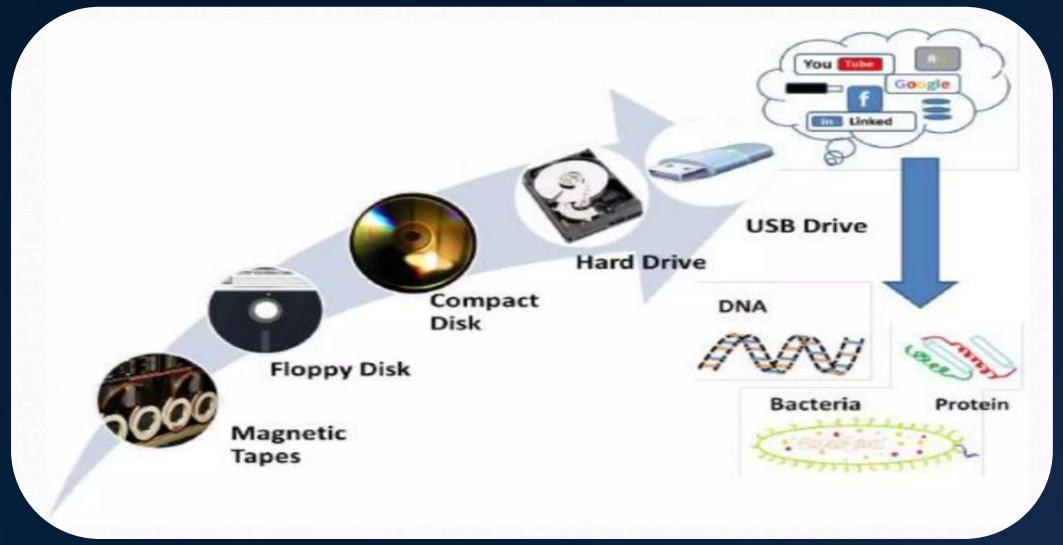
- Deoxyribonucleic acid (DNA) is the molecule that carries the genetic information for the development and functioning of an organism.
- The two DNA strands are called polynucleotides since they are composed of simpler monomer units called nucleotides.
- Each nucleotide is composed of one of four nitrogen-containing nucleobases - cytosine (C), guanine (G), adenine (A) or thymine (T).



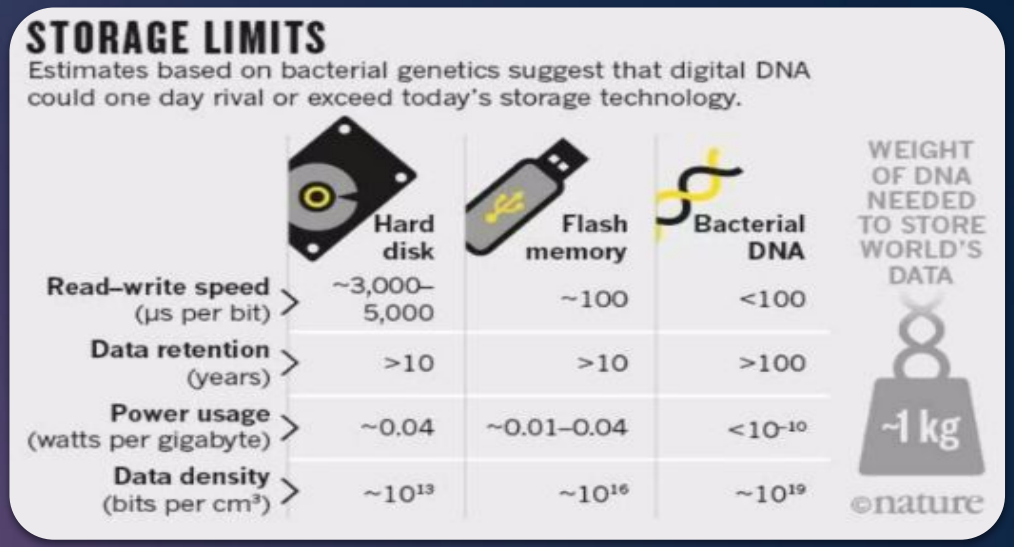
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.healthline.com%2Fhealth%2Fwhat-is-dna&psig=AOvVaw0L-FCjgJhNtN_Qc68VlfX6&ust=1668779266725000&source=images&cd=vfe&ved=0CBAQjRxqFwoTCJCyg66ttfsCFQAAAAAdAAAAABAE

DNA STORAGE SYSTEM

- Finding the ways to store data efficiently and economically is getting more and more harder.
- The exotic solution can be archiving information in DNA molecules.



https://www.researchgate.net/profile/Manish-Gupta-61/publication/277023595/figure/fig1/AS:294506877472784@1447227319886/Advancement-in-the-field-of-data-storage-devices-is-shown-here-New-paradigm-to-store.png



https://media.springernature.com/w300/springer-static/image/art%3A10.1038%2F537022a/MediaObjects/41586_2016_Article_BF537022a_Figc_HTML.jpg

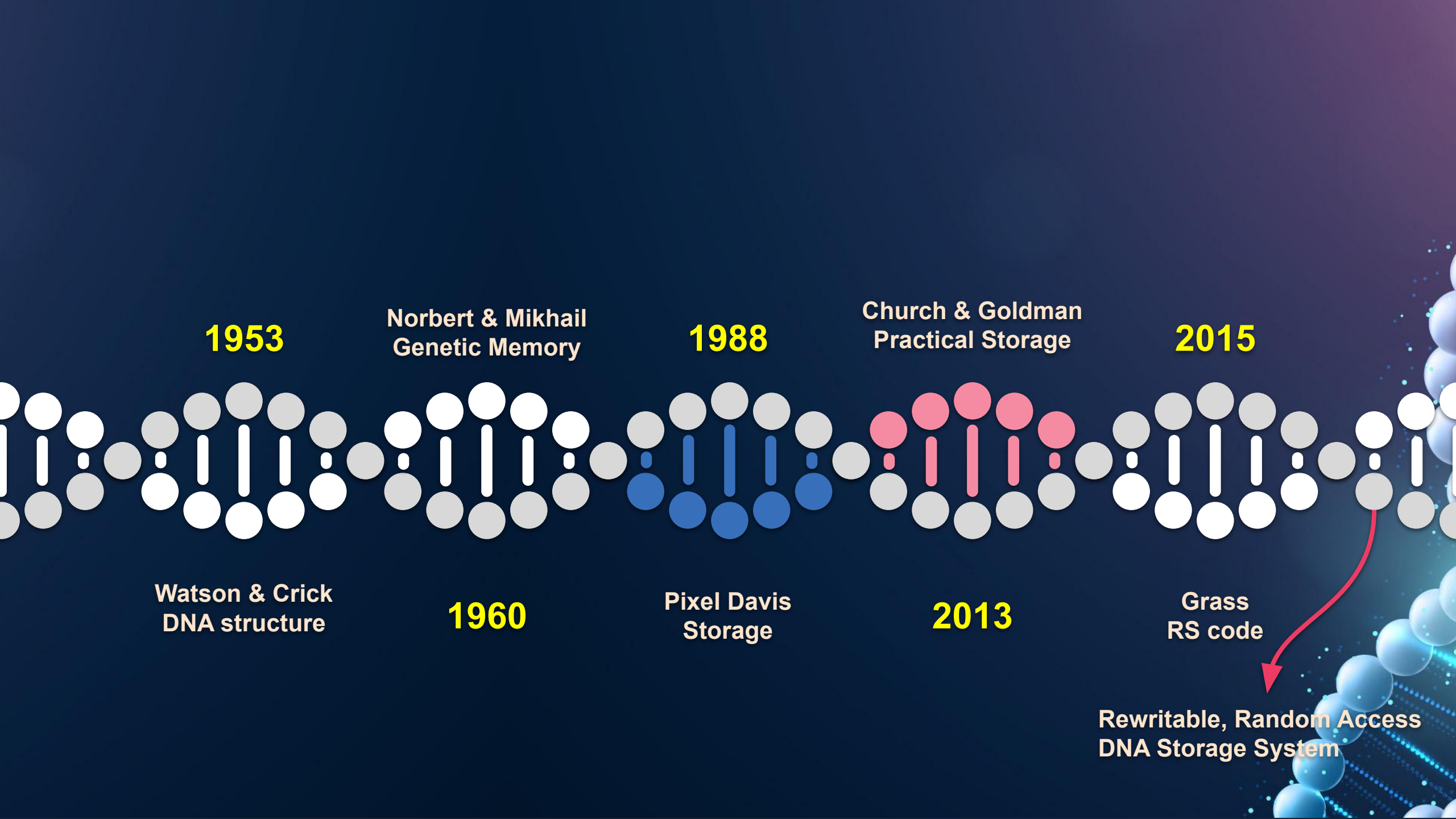
- DNA digital data storage is the process of transcoding of binary data to strands of DNA and vice versa.
- It has recently been announced that 1 gramme of DNA can contain 215 petabytes (215 million gigabytes).

2

Literature Review

RELATED WORK

Several studies have described designs for archival DNA-based storage in response to the growing need for enormous data repositories.


1953
Watson & Crick
DNA structure
Norbert & Mikhail
Genetic Memory
1960
1988
Pixel Davis
Storage
Church & Goldman
Practical Storage
2013
2015
Grass
RS code
Rewritable, Random Access
DNA Storage System

TRADITIONAL CODING SCHEMES

Single parity-check Coding

Compression of DATA

Huffman Coding

Removal of repeated consecutive bases in DNA String

Differential Coding

Addition of Controlled Redundancy which helps in reducing assembly errors

ISSUES

Both differential codes and Huffman codes are fixed-to-variable length compressors which could cause catastrophic error propagation when there is sequencing noise.

Single parity repetition encoding and differential encoding, sometimes known as RS codes, are unsuccessful against long substrings with a lot of GC content and other error-prone sequence patterns.

- The reason to reconstruct the whole text is to read or recover the information encoded even in a few bases is an even more significant issue.
- Additionally, all the current designs doesn’t support rewrite operation.

- The first restriction is a major drawback because it typically necessitates accommodating access to particular data parts.
- The second restriction prevents the use architectures that require remembering the history of edits, storing frequently updated data, and less frequent data editing.

SOLUTION

- We develop a re-writable and random-access DNA based storage architecture, with built-in error correction used for selective information access and encoding.
- The Addresses contained in the DNA strings are uncorrelated to each other.
- Encoding is accomplished by concatenating correctly terminated prefixes (of the addresses). Prefix encoding format allows block to be rewritten later.
- Rewriting is done using OE-PCR and gBlock DNA editing procedures.

3

About Our System

System Description

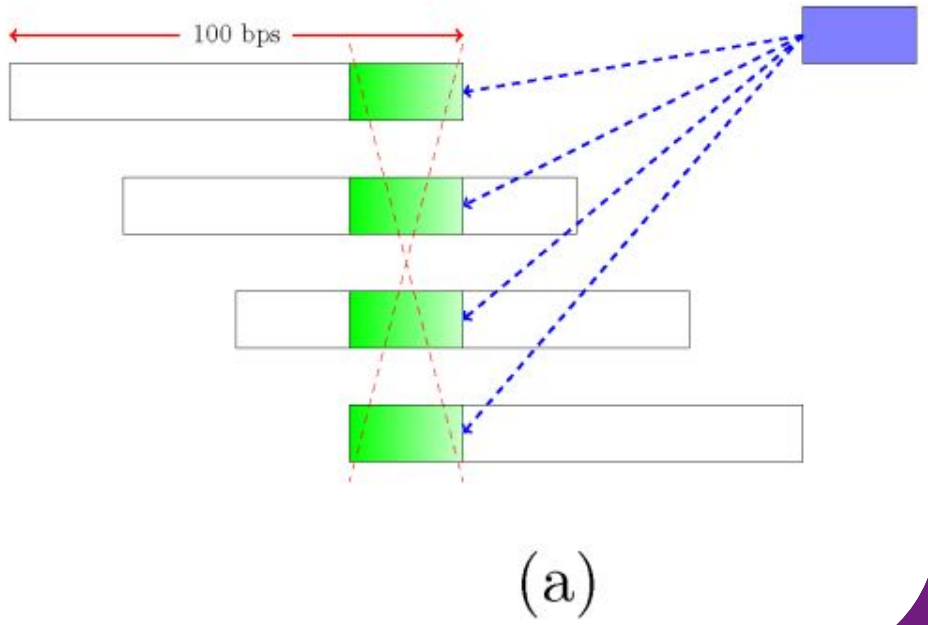
Our system uses severely scoped coding methods of editing DNA structures.

The system provides selective information access and encoding with built-in error-correction capabilities.

The addresses are intended to be mutually uncorrelated to be independent from each other.

This encoding technique represents a unique variation on prefix-synchronized coding.

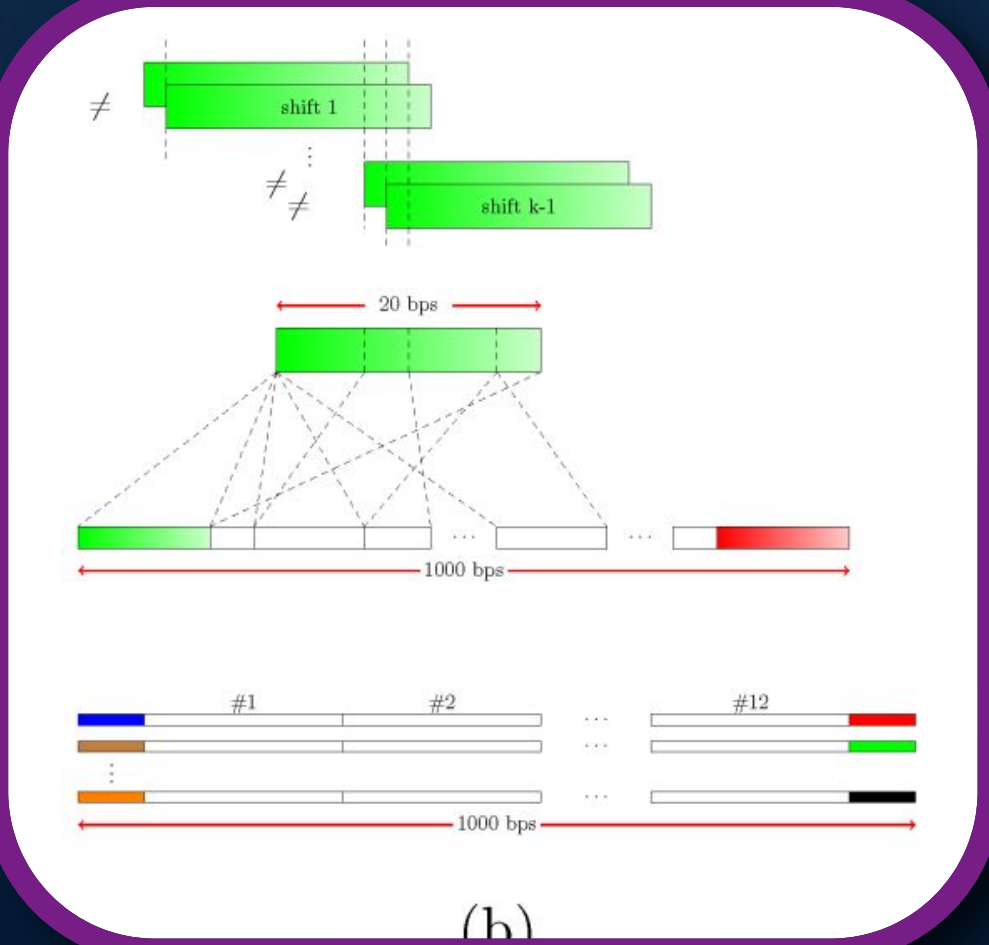
The system ensures data reliability, sensitivity and precision of retrieval while maintaining a high data storage capacity.



- The scheme's storage format consists DNA strings which covers encoded expressed text which is sampled of the size of 100 bps. Other than both ends, the fragments are overlapping on 75 bps, giving 4-fold coverage.

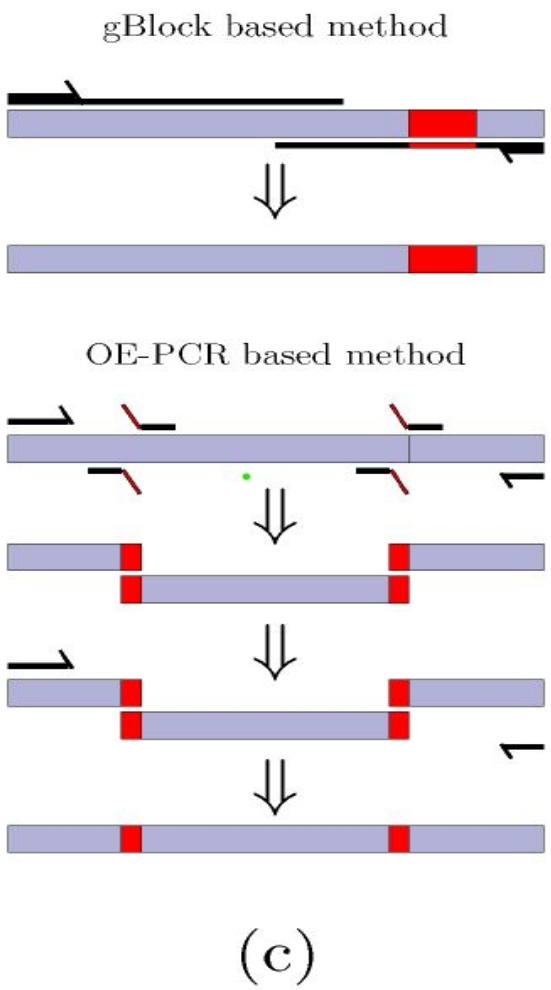
- If we want to rewrite a block, all its four fragments containing are needed to be selected and rewritten to record the new highlighted segment.

https://arxiv.org/pdf/1505.02199.pdf



https://arxiv.org/pdf/1505.02199.pdf

- The construction process of address sequence uses the idea of auto-correlation and cross - correlation.
- Mutually uncorrelated addresses are chosen and address of size 20 bps are appended on left and right side.



https://arxiv.org/pdf/1505.02199.pdf

- For content rewriting gBlock Method is used for short rewrites and OE-PCR method is used for sequential rewriting of longer blocks. OE-PCR is also cost-efficient.

COMMA FREE CODES

It is a block code where no concatenation of codewords that contains a valid codeword as substring.

It allows simple, yet efficient, synchronisation protocols.

A ‘drawback’ that in order to verify a string whether is a codeword or not we have to perform exhaustive search over set of sequences.

Overcoming that drawback by prefix synchronised codes

⇒ A special family of comma free codes.

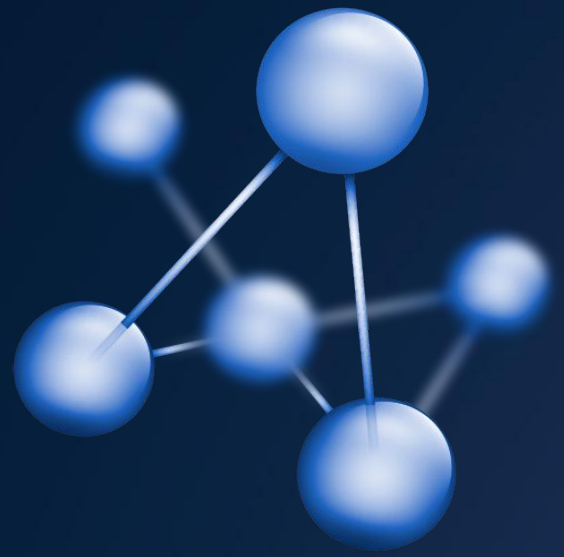
⇒ It has the form,

$$L = P + C$$

where, address prefix $P = p_1 p_2 \ldots p_n$

constrained sequence $C = c_1 c_2 \ldots c_s$

$$\Rightarrow L = p_1 p_2 \ldots p_n c_1 c_2 \ldots c_s$$



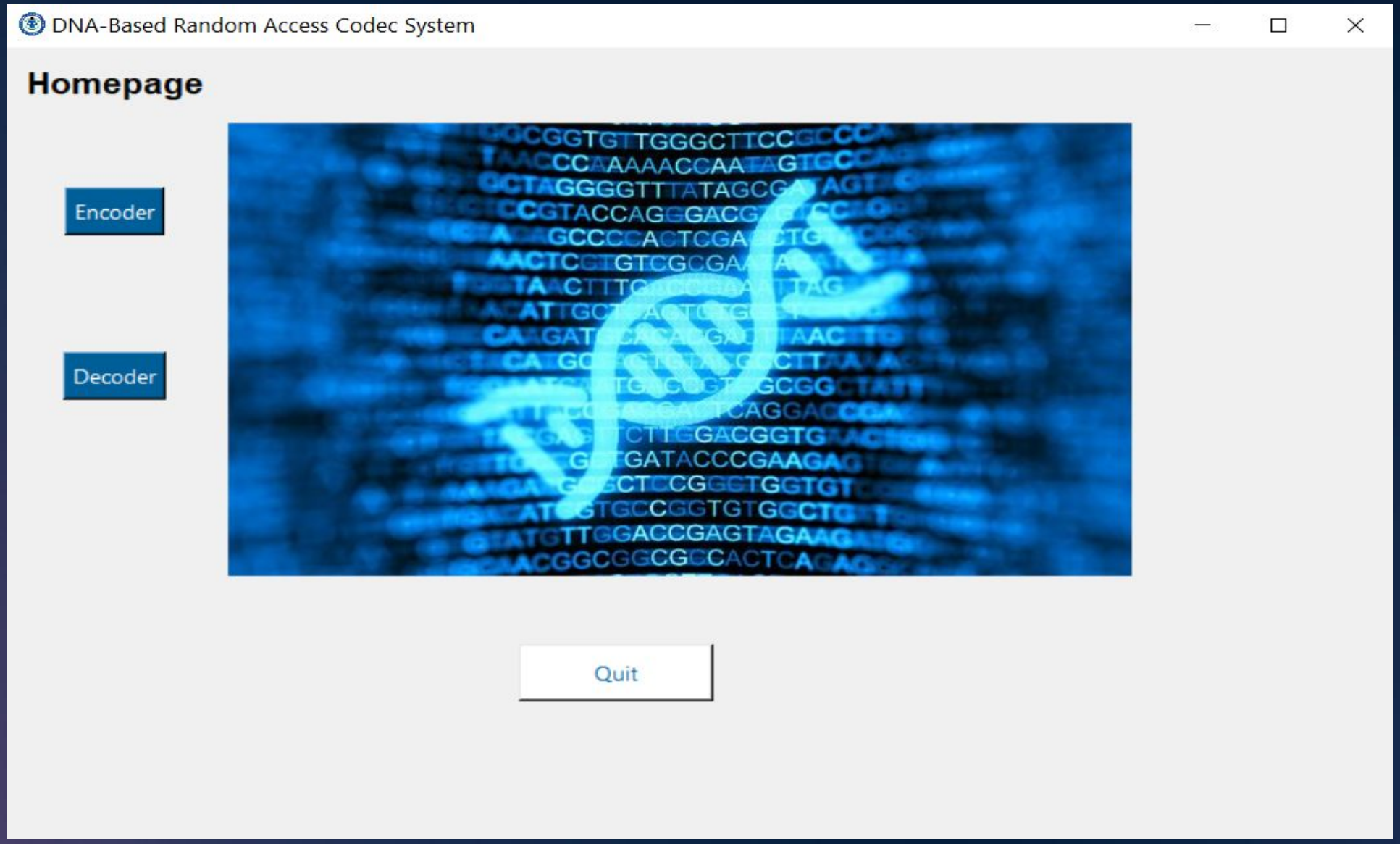
PREFIX SYNCHRONIZED ENCODING – DECODING

$X = \texttt{EncodePSC}(P, \ell, x)$
return $P\texttt{CodePSC}(P, \ell, x)$;

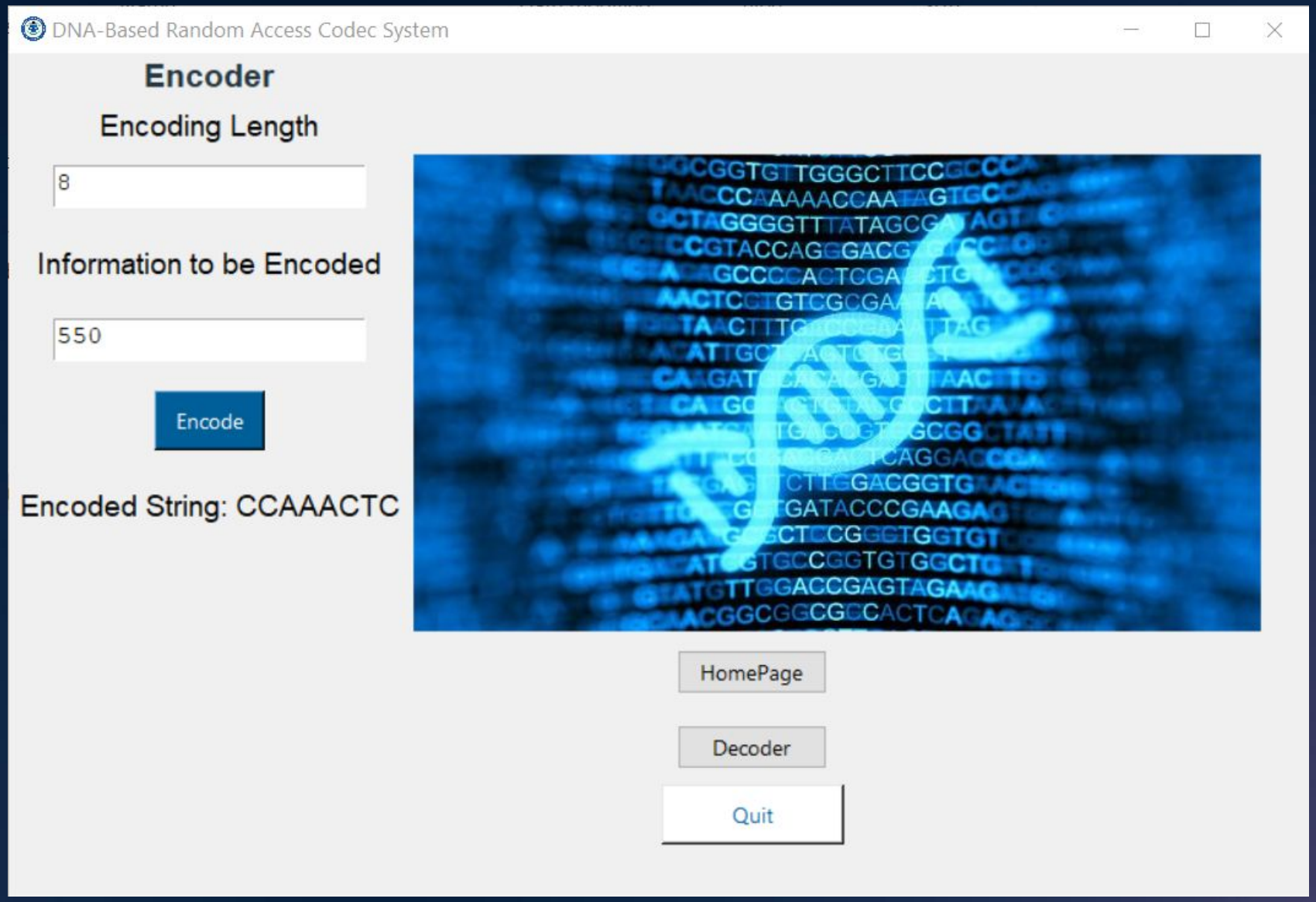
$X = \texttt{CodePSC}(P, \ell, x)$
begin
$n = \text{length}(P)$;
if $(\ell \geq n)$
$t := 1$;
$y := x$;
while $\left(y \geq \left|\bar{P}_t\right| G_{n,\ell-t}\right)$
$y := y - \left|\bar{P}_t\right| G_{n,\ell-t}$;
$t++$;
end;
$a := \left\lfloor \frac{y}{G_{n,\ell-t}} \right\rfloor$;
$b := \text{mod}\ (y, G_{n,\ell-t})$;
return $P^{t-1}\bar{p}_{t,a+1}\texttt{CodePSC}(P, \ell - t, b)$;
else
return $\theta_\ell(y)$;
end;
end;

$x = \texttt{DecodePSC}(P, X)$
begin
$n = \text{length}(P)$;
$\ell = \text{length}\ (X)$;
$X = X_1 X_2 \ldots X_\ell$;
if $(\ell < n)$
return $\theta^{-1}(X)$;
else
find$\left(s, t \text{ such that } P^{t-1}\bar{p}_{t,s} = X_1 \ldots X_t\right)$;
return $\left(\sum_{i=1}^{t-1} \left|\bar{P}_i\right| G_{n,\ell-i}\right) + (s-1) G_{n,\ell-t} + \texttt{DecodePSC}(P, X_{t+1} \ldots X_\ell)$;
end;
end;

Software Homepage



Encoder


DNA-Based Random Access Codec System
Encoder
Encoding Length
8
Information to be Encoded
550
Encode
Encoded String: CCAAACTC
HomePage
Decoder
Quit

Decoder


DNA-Based Random Access Codec System
Decoder
String to be Decoded
CCAAACTC
Decode
Decoded Address: 550
Encoder
HomePage
Quit

Contribution

- Jemish Variya - 20%
- Nisarg Nampurkar - 20%
- Harikrishna Patel - 20%
- Raj Patel - 20%
- Ashray Kothari - 20%

REFERENCES

[1] E. Gilbert. Synchronization of binary messages.
IRE Transactions on Information Theory, 6(4):470–477, 1960.

[2] Maya Levy and Eitan Yaakobi. Mutually uncorrelated codes for dna storage.
In 2017 IEEE International Symposium on Information Theory (ISIT), pages 3115–3119, 2017.

[3] S. M. Hossein Tabatabaei Yazdi, Yongbo Yuan, Jian Ma, Huimin Zhao, and Olgica Milenkovic. A rewritable, random-access dna-based storage system.
Scientific Reports, 5, 2015.

THANK YOU!