DNA Data Storage: Fundamentals and Challenges

Graduate Research Seminar – Bilkent University

Presenter: Qiyi Yao

Mar. 13, 2025

CONTENT

Catalogue

01 Background & Motivation

Why DNA?

02 DNA Storage Pipeline

Writing and Reading

03 DNA Storage Channel

Noisy Shuffling-Sampling Channel

04 Challenges

Coding, System Design, etc.

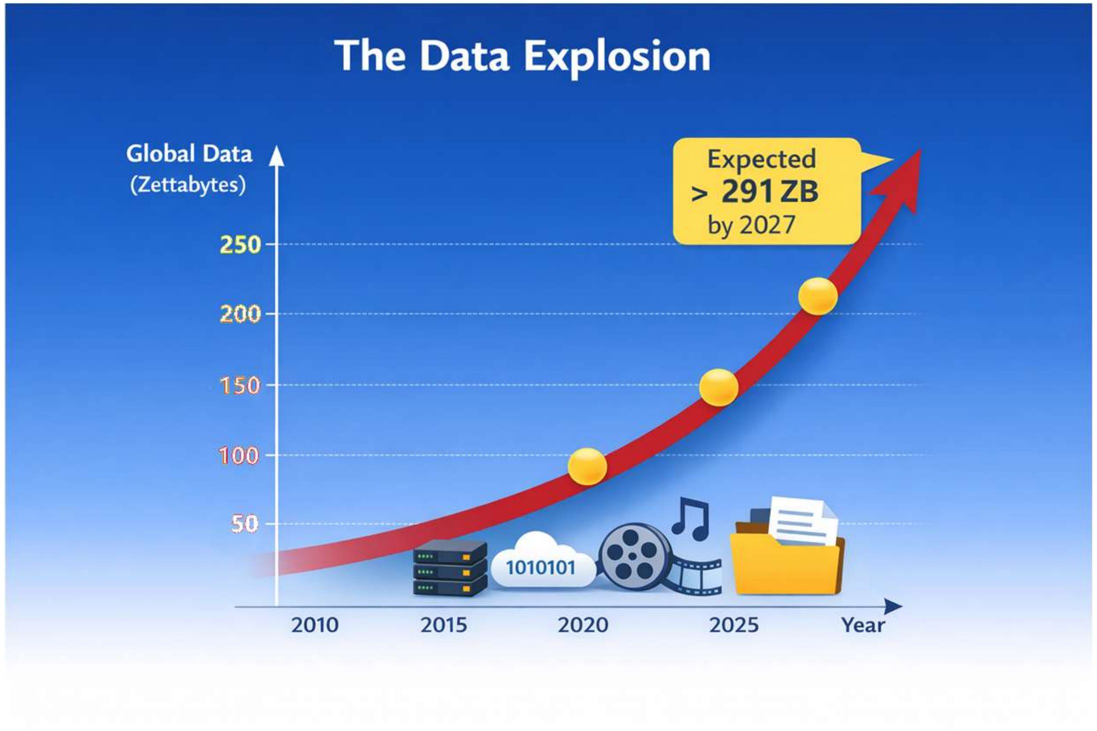
1

Background & Motivation

One kilogram of DNA could store the world’s data.

The Data Explosion

Background & Motivation



01 Global data are growing exponentially: by 2027, the expected global data will be more than 291ZB.

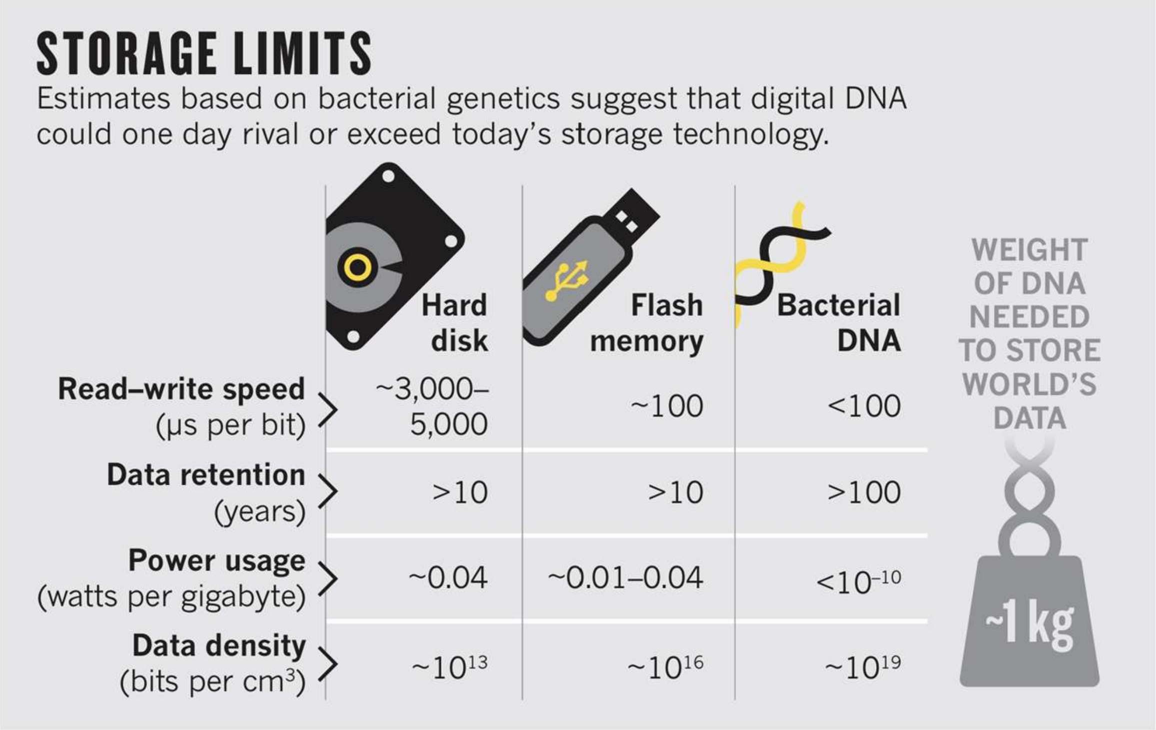
02 Long-term archival storage is expensive: data centers consume large energy and need high maintenance costs.

03 Traditional storage media have limitations: short lifespan, low storage density, etc.

IDC Worldwide Global DataSphere Forecast, 2023–2027.

Why DNA?

Background & Motivation



STORAGE LIMITS

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.

	Hard disk	Flash memory	Bacterial DNA
Read–write speed (μs per bit)	~3,000–5,000	~100	<100
Data retention (years)	>10	>10	>100
Power usage (watts per gigabyte)	~0.04	~0.01–0.04	$<10^{-10}$
Data density (bits per cm³)	$\sim10^{13}$	$\sim10^{16}$	$\sim10^{19}$

WEIGHT OF DNA NEEDED TO STORE WORLD'S DATA

~1 kg

01 Long lifespan

02 Low power usage

03 High Storage Density

Extance, Andy. "How DNA could store all the world's data." Nature 537.7618 (2016).

DNA Basics

Background & Motivation

01

Nucleotide

Sugar-phosphate backbone and **nucleotide bases: A, C, G, T**

02

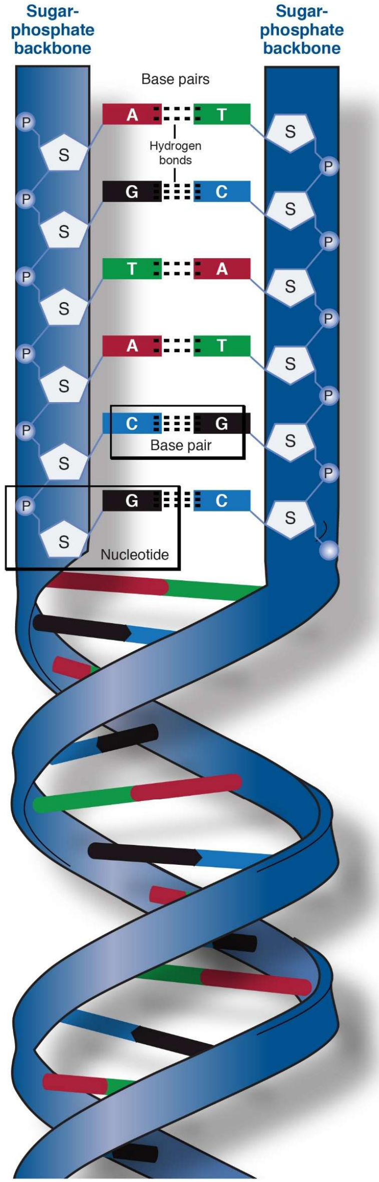
Double Helix

Two polynucleotide chains that coil around each other to form a double helix.

03

Base Pairs

Each nucleotide base on one polynucleotide chain forms hydrogen bonds with another nucleotide base on the other chain. Base pairing rules: **A with T (two hydrogen bonds); G with C (three hydrogen bonds)**.



2

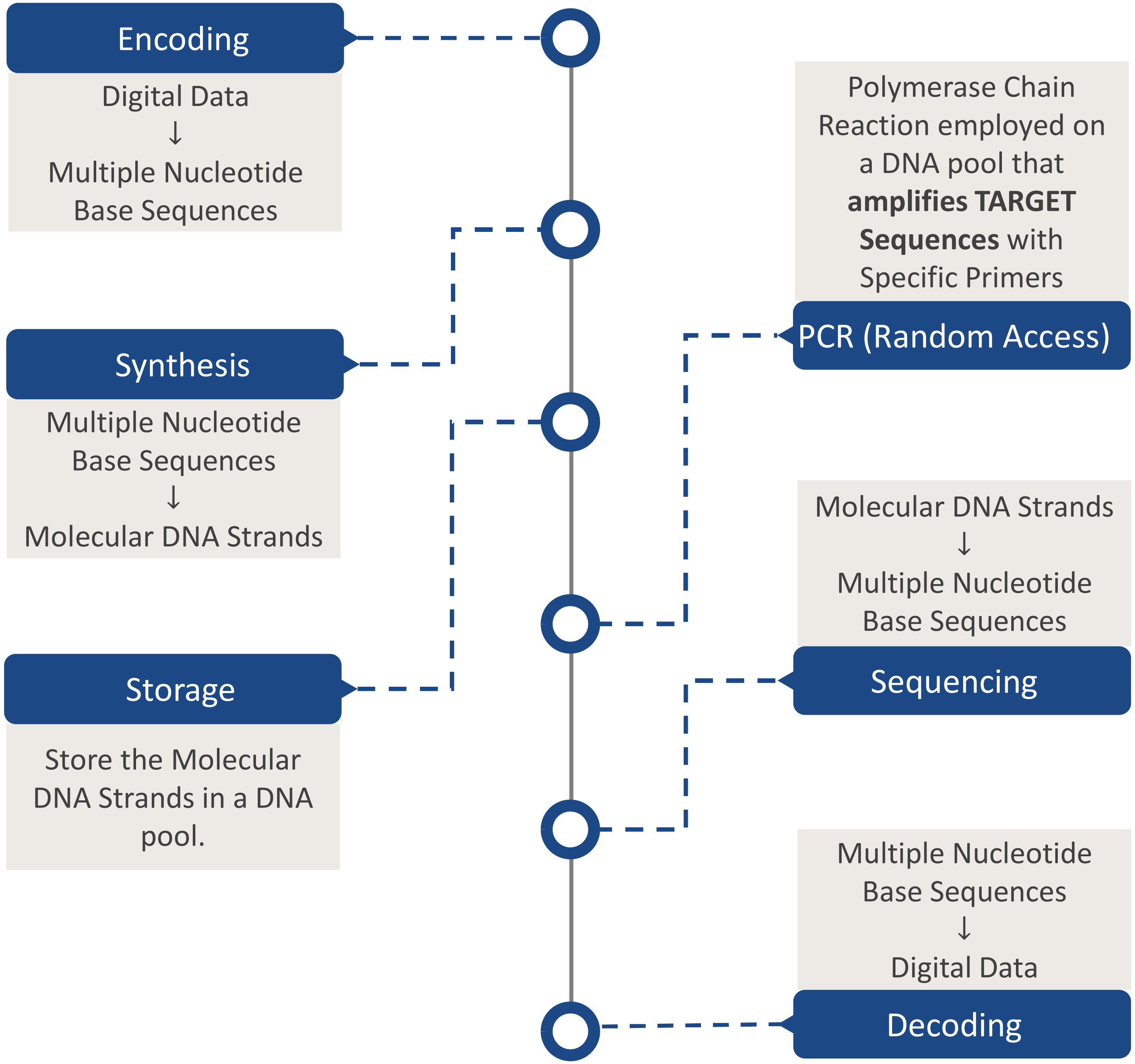
DNA Storage Pipeline

What are the components of a DNA storage system?

Overview

DNA Storage Pipeline

Writing



Encoding

Digital Data
↓
Multiple Nucleotide Base Sequences

Synthesis

Multiple Nucleotide Base Sequences
↓
Molecular DNA Strands

Storage

Store the Molecular DNA Strands in a DNA pool.

PCR (Random Access)

Polymerase Chain Reaction employed on a DNA pool that **amplifies TARGET Sequences** with Specific Primers

Sequencing

Molecular DNA Strands
↓
Multiple Nucleotide Base Sequences

Decoding

Multiple Nucleotide Base Sequences
↓
Digital Data

Reading

Encoding & Decoding

DNA Storage Pipeline

How DNA Encodes Information?

A simple example:

00 → A
01 → C
10 → G
11 → T

01 01 10 10 00 01 11

↓

C C G G A C T

Does this native code work?

Encoding & Decoding

DNA Storage Pipeline

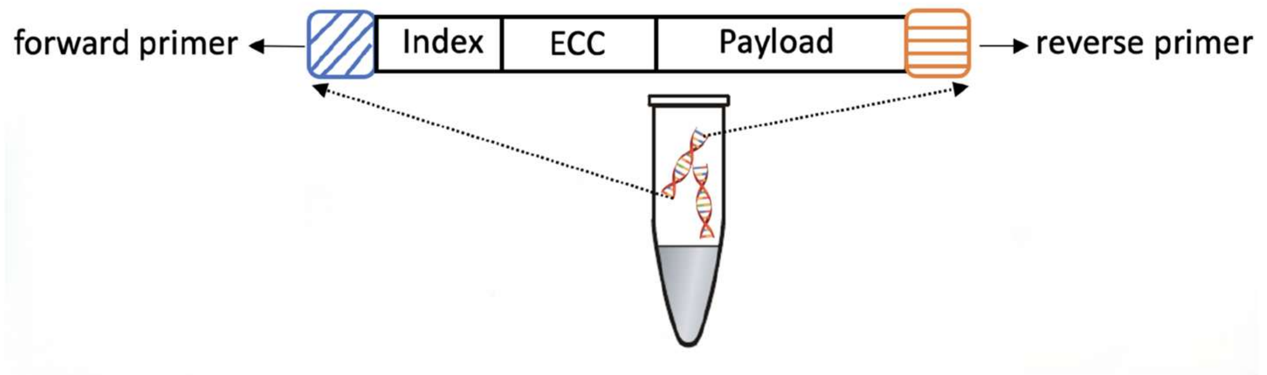
How DNA Encodes Information?

A simple example:

00 → A
01 → C
10 → G
11 → T

01 01 10 10 00 01 11
↓
C C G G A C T

Does this native code work?



No!

Sensintaffar, Alex, et al. "Advancing archival data storage: The promises and challenges of dna storage system." ACM Transactions on Storage 21.3 (2025): 1-34.

Synthesis & Storage

DNA Storage Pipeline

DNA Synthesis

General approach:

- Adding nucleotides as base pairs (bp) one at a time until a subset of the DNA strand is created.
- Combine these subsets into a single DNA strand, which can be later duplicated to produce additional copies.

Problem: DNA Synthesis introduces Insertion/Deletion/Substitution (IDS) errors.

Problem: The synthesis error rate increases exponentially as the strand length increases.

Synthesis & Storage

DNA Storage Pipeline

DNA Synthesis

General approach:

- Adding nucleotides as base pairs (bp) one at a time until a subset of the DNA strand is created.
- Combine these subsets into a single DNA strand, which can be later duplicated to produce additional copies.

Problem: DNA Synthesis introduces Insertion/Deletion/Substitution (IDS) errors.

Problem: The synthesis error rate increases exponentially as the strand length increases.

DNA Storage

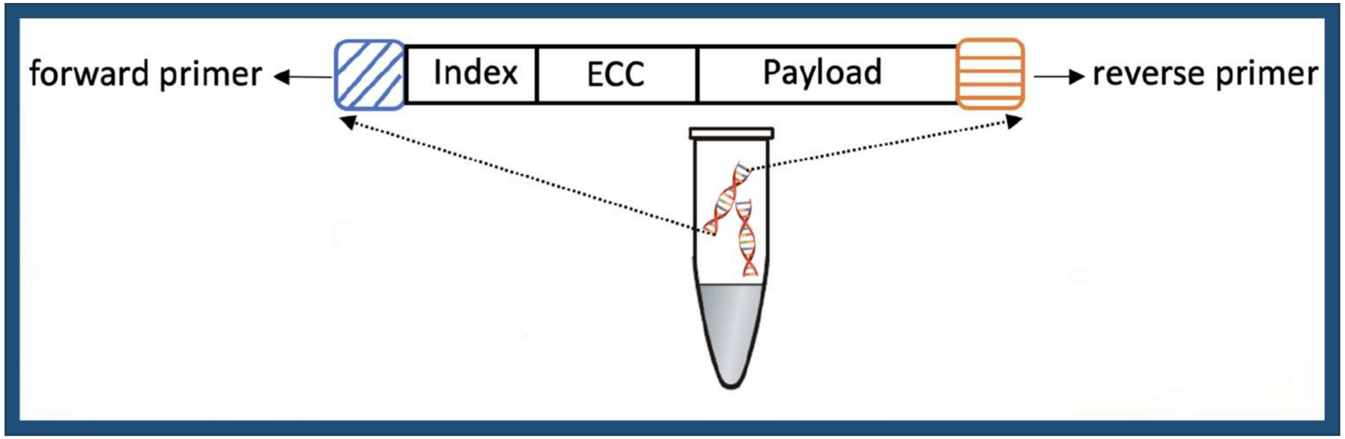
- Multiple short DNA strands are synthesized and stored in a DNA pool (a tube or other container).
- For a certain DNA strand inside a DNA pool, it's not able to directly access it: random sampling over the pool is required.
- The strands can be considered shuffled.

Problem (Random Access): How to directly access the strands that we want to read.

Problem: Since the DNA strands are shuffled, how to restore the order of the stored information?

PCR-Based Random Access

DNA Storage Pipeline



Polymerase Chain Reaction (PCR)

Repeatable technique used to amplify (duplicate) targeted DNA strands with a specifically assigned primer pair. In each iteration of PCR, the number of strands with the specific primer pair doubles.

Random Access

For a certain set of strands with a shared primer pair, through multiple iterations of PCR that amplify the set of strands, random sampling will result in the target strands with high probability.

Organick, Lee, et al. "Random access in large-scale DNA data storage." Nature biotechnology 36.3 (2018): 242-248.

Three Generations of Sequencers

First Generation

E.g. Chain-Terminating Inhibitors

Could accurately read long DNA strands and provide high-accuracy, high-quality sequences with high costs, long sequencing time, and labor-intensive procedures.

Second Generation

E.g. Reversible Terminator Chemistry

Faster than the first-generation sequencers with reduced cost and labor, but it has short read lengths (hundreds) and difficulty in reading long homopolymers.

Third Generation

Nanopore Sequencing

Significantly faster than the second-generation sequencers with much longer read lengths (thousands) but has higher error rates.

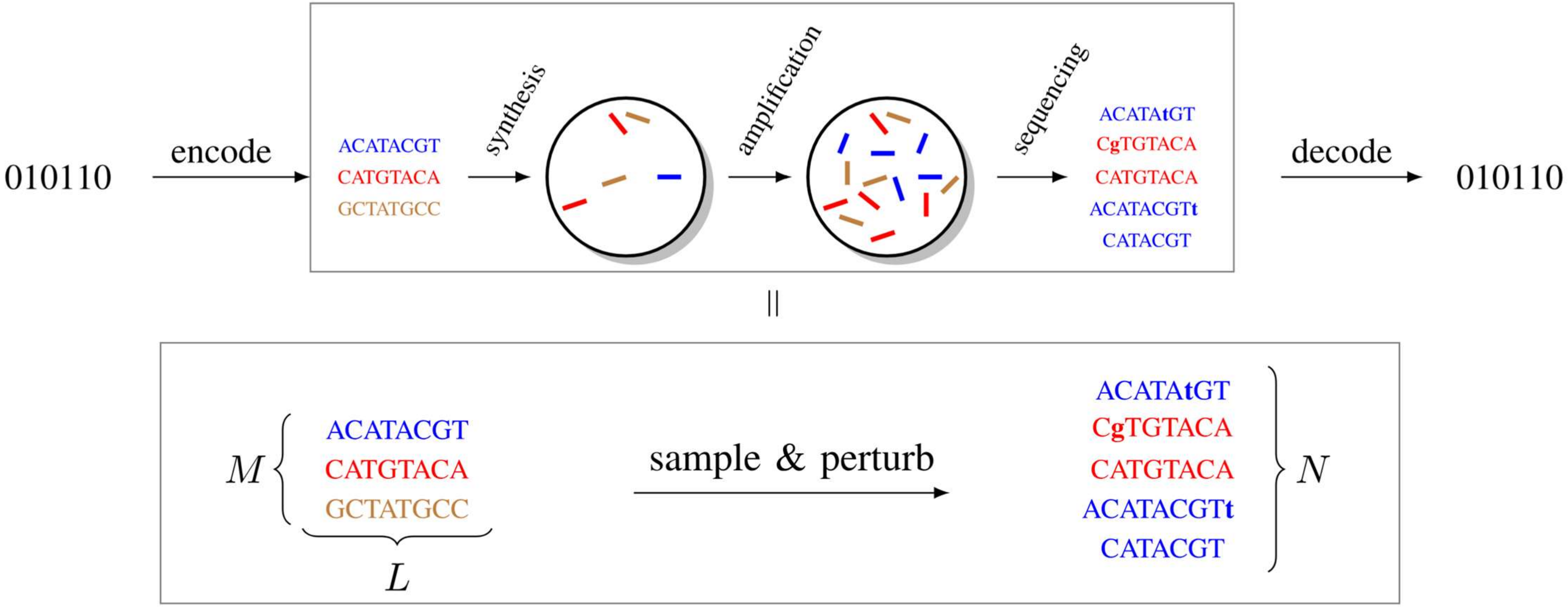
3

DNA Storage Channel

Modeled as noisy shuffling-sampling channel.

The Noisy Shuffling-Sampling Channel

DNA Storage Channel



Shomorony, Ilan, and Reinhard Heckel. "DNA-based storage: Models and fundamental limits." IEEE Transactions on Information Theory 67.6 (2021): 3675-3689.

DNA Storage Codes

DNA Storage Channel

Reducing Channel Noise

Biological Constraints:

- **GC-Content**
- **Homopolymer Run Length** Constraint

Error Correction

Codes and coding methods that correct insertion/deletion/substitution (IDS) errors

(Internal) Indexing

- **Order Restoration** within strands sharing the same primer pair
- Identification of certain strands

Primer Library Design

Large Hamming Distance between each two distinct primers in the primer library

Biological Constraints & ECC

DNA Storage Channel

Homopolymer Run Length: Sub-sequences with identical nucleotides occurring consecutively experience a higher error rate compared to sub-sequences without homopolymers.

GC Content: the proportion of G and C nucleotides in a DNA strand directly affects the thermal stability of the DNA strand, which in turn influences the strand’ s overall lifetime.

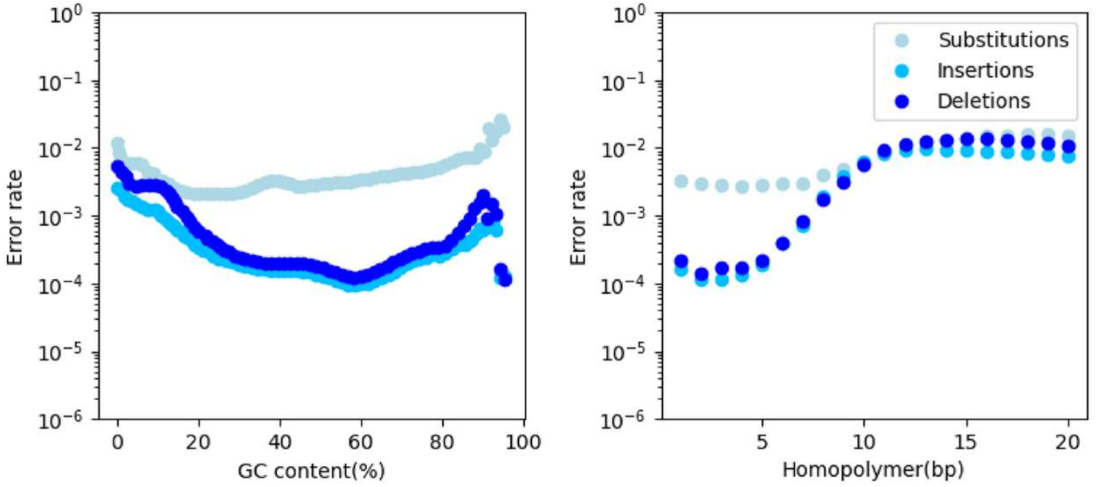
Reducing Channel Noise

Biological Constraints:

- **GC-Content**
- **Homopolymer Run Length** Constraint

Error Correction

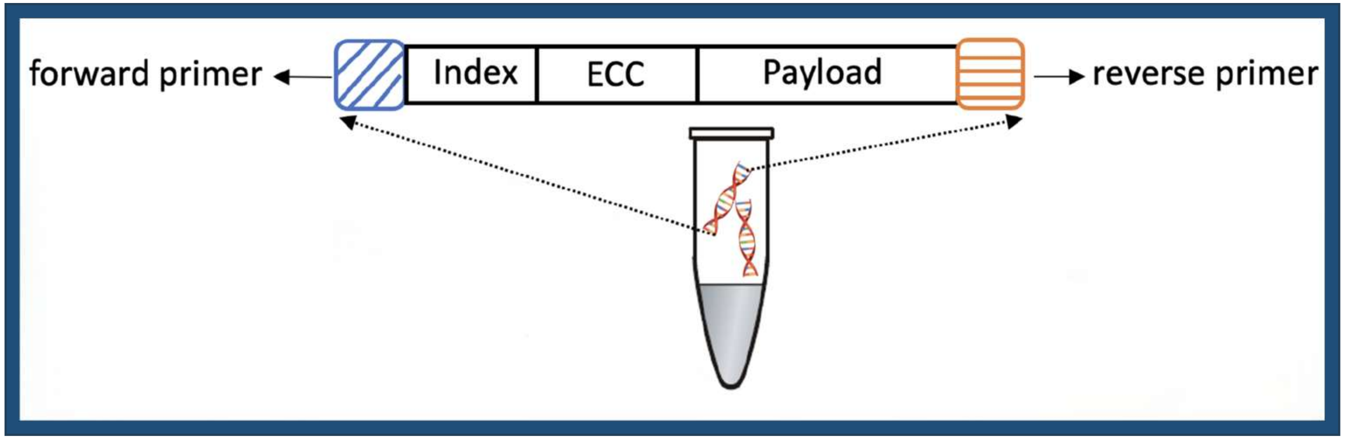
Codes and coding methods that correct insertion/deletion/substitution (IDS) errors



Ross, Michael G., et al. "Characterizing and measuring bias in sequence data." Genome biology 14.5 (2013): R51.

Indexing & Primer Library

DNA Storage Channel



(Internal) Indexing

- **Order Restoration** within strands sharing the same primer pair
- Identification of certain strands

Explicit Indexing: Indexes are encoded and appended to the DNA strands.

Implicit Indexing: Concatenated coded index where the indexes are expressed by the sub-codes of the inner code.

Primer Library Design

Large Hamming Distance between each two distinct primers in the primer library

If any two primers are too similar (with relatively small Hamming distance), then wrong DNA strands could be amplified during the PCR process.

4

Challenges

Countering system-level errors, maximizing storage capacity, etc.

Countering Errors

Challenges

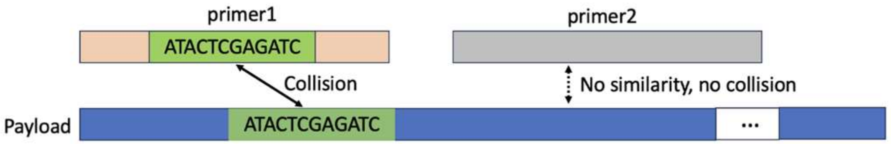
Primer Library Requirement

End Constraint

The last five nucleotides of the DNA strand cannot have more than three GC nucleotides.

Primer-payload Collision

A primer and a payload share a pair of nearly identical subsequences.



Resulting in amplification of irrelevant payload.

System-Level Code Requirement

Inter-complementary

The occurrence of any two distinct sequences that are complementary to each other.

Intra-complementary

The occurrence of any two subsequences, within a sequence, reverse complementary to each other.

E.g. ATGA-TCAT → ATGA
TACT

ATGA
TACT

Instead of

ATGATCAT
TACTAGTA

Maximizing Storage Capacity

Challenges

Storage Capacity

Encoding Density

Coding rate (bits per nucleotide)

New code designs to increase Encoding Density

DNA Payload Length

Due to practical limitations, DNA strands cannot be too long.

How to reduce Indexing Overhead

Parallel Factor

The number of unique DNA strands that share one primer pair and can be sequenced out together.

Number of Usable Primers

How many primers can be safely used for PCR-based random access in a single DNA pool.

Dealing with Primer-Payload Collision

Speeding up Reading

Challenges

Leading Causes for Long Read Latency

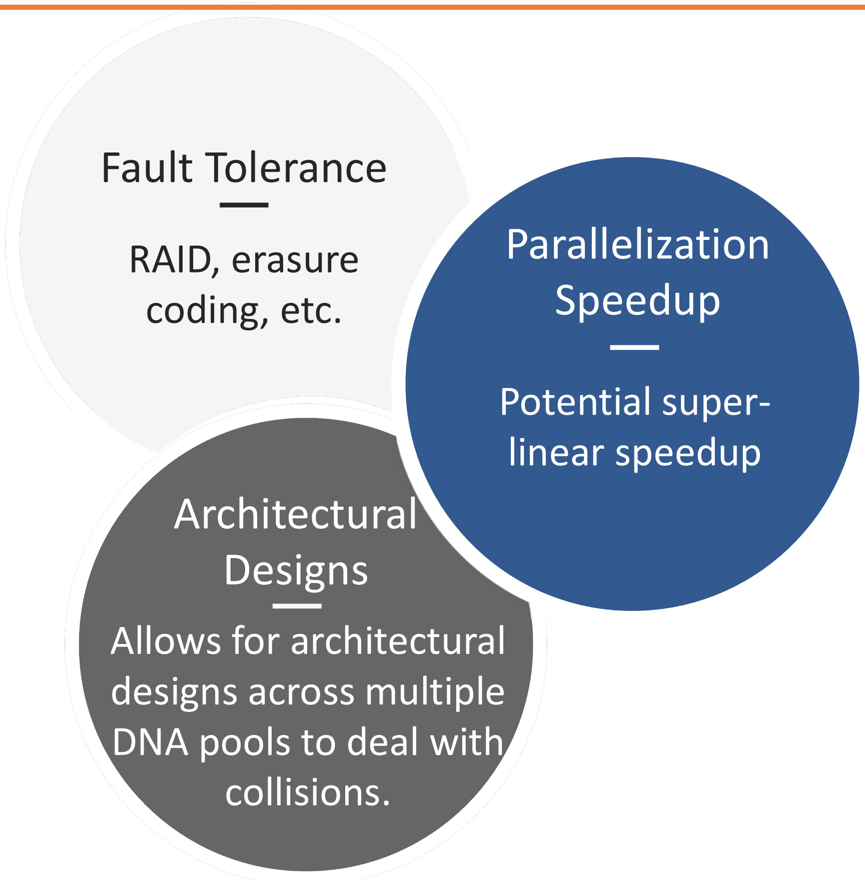
Large number of rounds of the PCR process

The need for repeated reads following the random sampling nature of DNA storage

Parallelization & Distributed Storage

Leading Causes for Long Read Latency

- Large number of rounds of the PCR process
- The need for repeated reads following the random sampling nature of DNA storage



Fault Tolerance

RAID, erasure coding, etc.

Parallelization Speedup

Potential super-linear speedup

Architectural Designs

Allows for architectural designs across multiple DNA pools to deal with collisions.

Benefits of Parallelization and Distributed Storage

Writing & Update

Challenges

Speeding up Writing

Approaches:

- Motifs: Prefabricated DNA Sequences (Using prefabricated motifs allows for modular designs and DNA strands no longer need to be built from scratch.)
- Parallelization

DNA Storage Update

Difficulty:

Massive Duplications of DNA Strands:

- DNA strands are duplicated many times due to PCR
- Updating a certain DNA strand or a certain set of DNA strands requires updating the copies correctly as well

Approach:

- Design **deduplication** systems for DNA data storage.

Q & A

Thank You!