DNA-STORALATOR: END-TO-END DNA STORAGE SIMULATOR

Gadi Chaykin, Nili Furman, Omer Sabary, Dvir Ben Shabat, Eitan Yaakobi

Non-Volatile Memory Workshop 2022
INTRODUCTION

Writing Information

1. Encoding
   - Binary data file
   - 001010
   - 110100
   - 010011
   - 111011
   - 110010
   - ACGGACA
   - AACGCGA
   - CCGATA
   - TTAGTACA
   - GGACTCA

2. Synthesis
   - Encoded data
   - Storage container
   - Multiple copies for each DNA strand
   - Errors in the stored strands

3. Sequencing
   - Noisy copies of the encoded data
   - CCGATAG
   - GGGACTCA
   - TTAGTACA
   - ACGGACA
   - GGACCTAA

Generating Information

4. Decoding
   - ACGGACA
   - AACGCGA
   - CCGATAG
   - TTAGTACA
   - GGACTCA

Created with BioRender
THE BIOLOGY RELATED STEPS of DNA STORAGE

Expensive and complicated, and therefore are not widely accessible to the community.
**PRICING**

- **Synthesis**
  - 100,000 200-base strands cost ≈ $10-15K (1MB = $3-5K)

- **Sequencing**
  - Illumina Hiseq
    - $2-3K for 200M strands
  - Oxford Nanopore Technologies MinION sequencer
    - $1000 for a single run (flow cell) to read 50M strands where each is 1000 bases.
Both synthesis and sequencing can cause errors.

- Deletions
- Insertions
- Substitution
ERRORS IN DNA - REASONS

INTRODUCTION

Synthesis
- Mostly for chemical reasons
- Each copy of a certain sequence has different errors

PCR
- Creates a bias - prefers one sequence over another

Sequencing
- Higher GC content affects sequencing error
- Presence of Homopolymers increases the error rate
TECHNOLOGY LIMITATIONS

- Synthesized strands are limited in their length (roughly up to 300 symbols).

- Multiple (thousand to millions) noisy copies per designed strands.

- The noisy copies are mixed and stored together.
MOTIVATION

• DNA storage includes insertion and deletion errors.
  • Much more complicated to correct, compared to classical storage media.
  • New coding schemes, algorithms, and techniques are required.

• Synthesis, PCR, and sequencing are complicated and expensive.
  • Therefore, not widely accessible to the community.
MESA: SIMULATION OF DNA SYNTHESIS, STORAGE, SEQUENCING AND PCR ERRORS

- A simulator for the processes of synthesis, PCR, storage, and sequencing errors.
- Includes detailed description of the errors involved and factors such as temperature, storage time, etc.

DEEPSIMULATOR: A DEEP SIMULATOR FOR DNA SEQUENCING

- Simulates nanopore sequencing: including the raw signal.
- Deep-learning-based tool.
NANOPORE SEQUENCING SIMULATOR FOR DNA DATA STORAGE

- Simulates the sequencing, storage and PCR processes.
- Based on a 2 years long experiment that evaluated how time effects the errors.

THE STORALATOR

A software tool that allows researchers from all fields to compare, study, and improve their coding techniques and algorithms with current state-of-the-art solutions.
A cross-platform software tool that simulates the processes of synthesis, PCR, sequencing and the algorithmic part of clustering and reconstruction of digital data in DNA molecules.

The tool can simulate the errors of the synthesis, PCR, and sequencing processes for the different available technologies.
DNA-STORALATOR: MODULES

- SOLQC: Error characterization
- Clustering
- Error Simulation: Synthesis, PCR, Sequencing
- Reconstruction
Sample from a DNA library → Error characterization (SOLQC) → Error statistics and cluster size distribution

Designed DNA strands → Synthesis and sequencing error simulation

Shuffled noisy reads → Clustering algorithms

Clustered noisy reads → Reconstruction algorithms

Estimation of the designed DNA strands
ERROR SIMULATION + CLUSTERING FLOW

INPUT

Design file

Selection of sequencing and synthesis technologies

Configure cluster sizes

Error simulation, based on analysis of prev. experiments.

Noisy reads: shuffled and clustered.

OUTPUT
SOLQC: SYNTHETIC OLIGO LIBRARY QUALITY CONTROL TOOL

**SOLQC**

- **Input**: synthetic DNA library (sequencing results + design file).
- Performs error characterization: error rates, and cluster size distribution.
- The Storalator’s error injection module is based on the analysis of previous wet experiments.
SOLQC RESULTS - EXAMPLE

ERROR RATES IN [1]

CLUSTER SIZE DISTRIBUTION [2]


SYNTHESIS AND SEQUENCING SIMULATION

- Simulates insertions, deletions, and substitutions which occur in the chemical processes.
- Provides a combination of technologies of synthesis and sequencing methods.
- Based on results from previous wet experiments.
- Allows user-defined error rates.
CLUSTER SIZE SIMULATION (PCR)

- Simulated by generating a different number of copies for every given designed strand.
- Number of copies in each cluster can be defined by:
  - Explicit definition.
  - Distribution - Probability density function of the cluster size distribution.
  - The default distribution is the skewed-normal distribution.
The goal: cluster the strands related to each other

**In house algorithms**
- Pseudo clustering algorithm – filter by threshold.
- Index-based clustering with options.

Implementation of previously published algorithm by Rashtchian et al. [1]:
- Min-hash based algorithm.

Output statistics
- Number of clusters generated.
- True-positive rates.
- False-negative rates.

---

RECONSTRUCTION ALGORITHMS

- The goal is to estimate the original strand from its noisy cluster.
- Input: clustered file of noisy reads.
- In house algorithms:
  - Linear time reconstruction algorithms
  - Dynamic-programming based algorithms
  - Trellis-based algorithms

USE CASE EXAMPLES

That can benefit the DNA storage community.
A DEVELOPMENT OF NEW CODING TECHNIQUES

• Analysis and comparison of coding techniques for DNA storage systems.

• Can be utilized to estimate the required error-correction capability of current/future DNA synthesis and sequencing methods.
USE CASE EXAMPLES

 Encode → Synthesis

 Clustering ↔ Sequencing

 Reconstruction → Decode
ENCODING/DECODING
BY

Chamaeleo: a robust library for DNA storage coding schemes
(2020)

Zhi Ping, Haoling Zhang, Shihong Chen, Qianlong Zhuang, Sha Joe Zhu, Yue She
RESULTS – BLAWAT ET AL.

USE CASE EXAMPLES

Hamming + RS  RS  No ECC

RESULTS – GOLDMAN ET AL.

1 deletion
2 insertions
1 substitution

1 deletion
1 substitution

DEVELOPMENT OF NEW ALGORITHMS FOR DNA STORAGE

The tool supplies a convenient way to compare new and existing algorithms for DNA storage systems.
EXPERIMENT DESIGNING

The storalator provides an efficient method to test new algorithms and coding techniques before performing expensive and time-consuming wet experiments.
What's next?
SOME OF OUR FUTURE PLANS

- Expand existing algorithms.
- Add new algorithms in all the different modules of the Storalator.
- Add new coding schemes for encoding/decoding.
- Expand the collaboration with users, researchers and developers.
- Present a new GUI for the tool with improved UX.
DO YOU HAVE MORE RESULTS?

You are welcome to share it with us, and we will happily analyze it using our tools!
SPECIAL THANKS TO

Gadi Chaykin
Nili Furman
Eitan Yaakobi
Dvir Ben-Shabat

For helping grow this project 😊
THANKS

Does anyone have any questions?