TECHNION
Israel Institute
of Technology

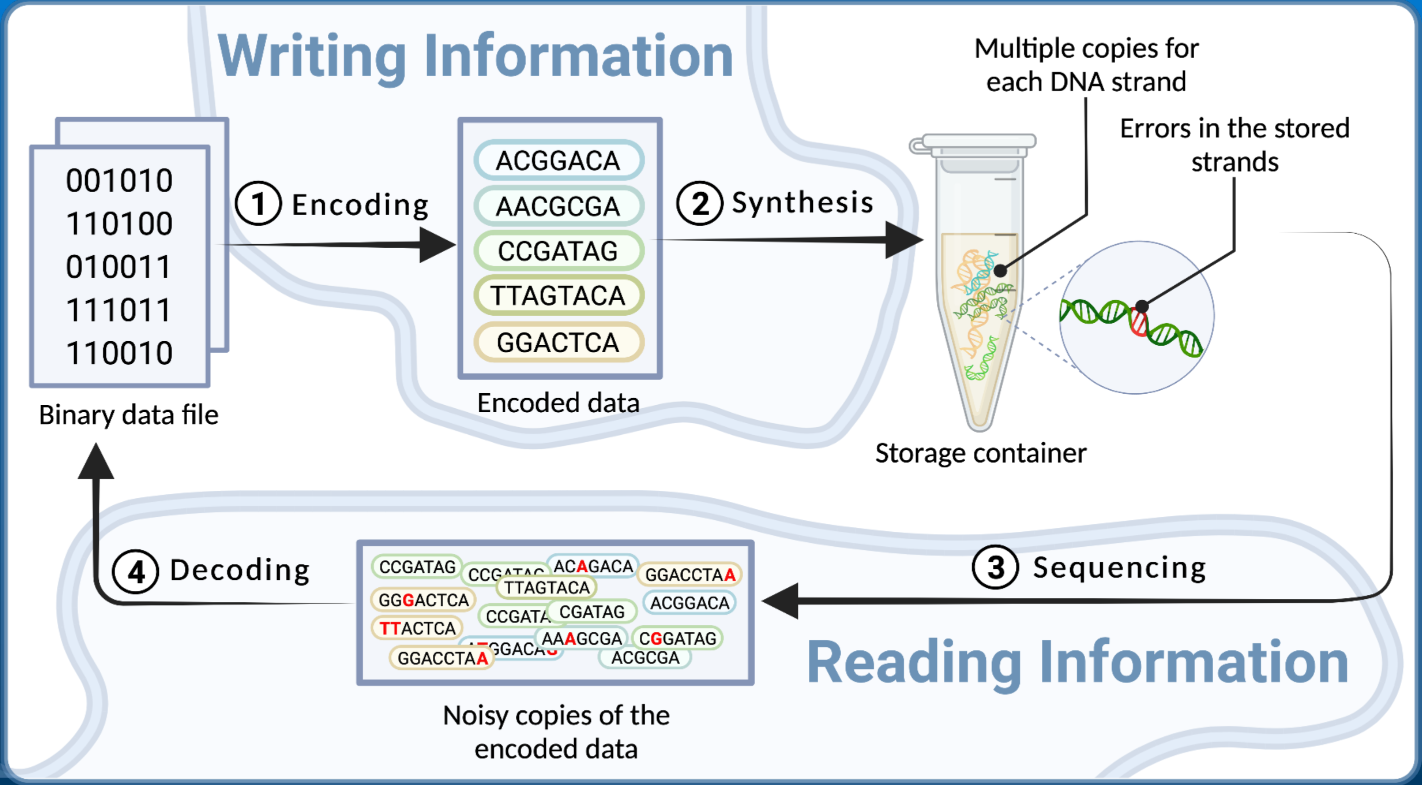
DNA-STORALATOR: END-TO-END DNA STORAGE SIMULATOR

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

Non-Volatile Memory Workshop 2022

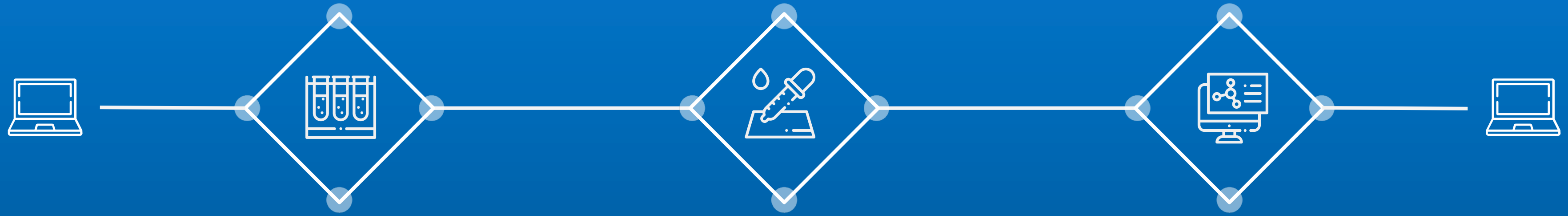
Speaker: Omer Sabary

INTRODUCTION



Created with BioRender

THE BIOLOGY RELATED STEPS of DNA STORAGE



SYNTHESIS

PCR

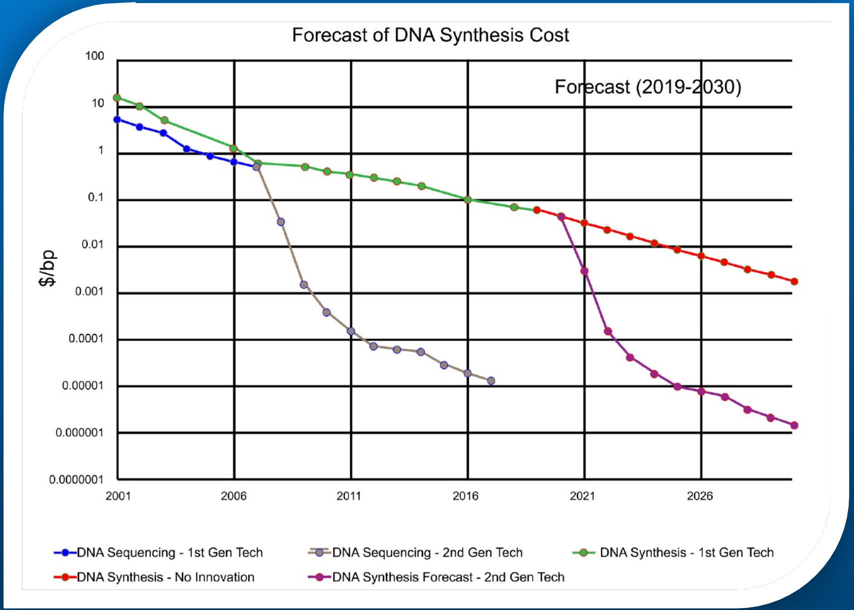
SEQUENCING

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)

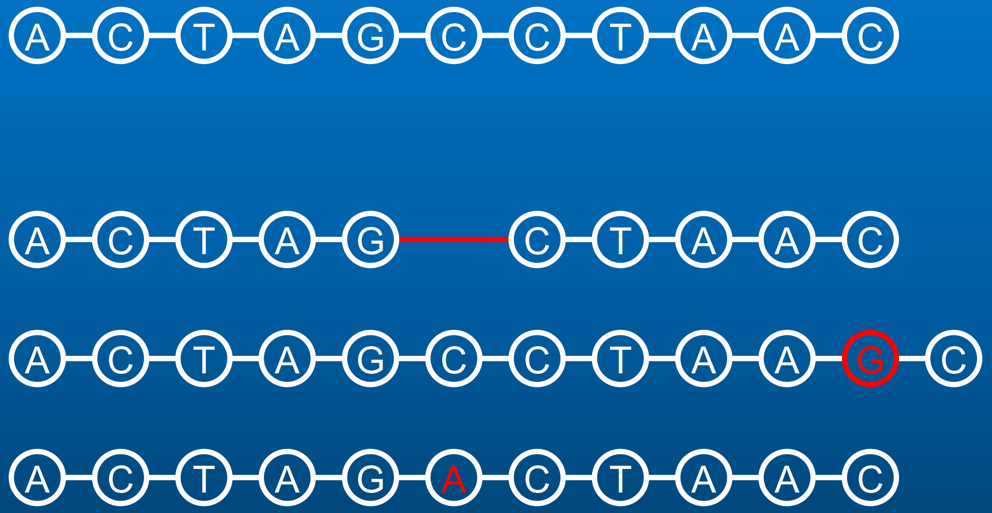
- **Sequencing**
  - Illumina Hiseq
    - $2-3K for 200M strands
  - Oxford Nanopore Technologies MinION sequencer
    - $1000 for a single run (flow cell) to read 50M strands where each is 1000 bases.



ERRORS IN DNA

Both synthesis and sequencing can cause errors.

- Deletions
- Insertions
- Substitution



ERRORS IN DNA - REASONS

Synthesis

Mostly for chemical reasons

Each copy of a certain sequence has different errors

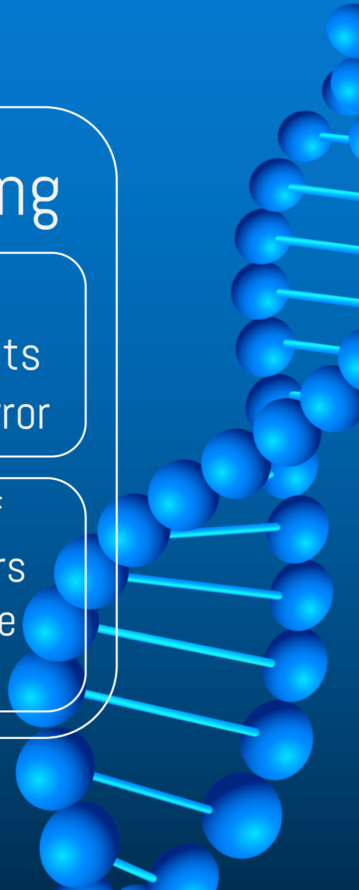
PCR

Creates a bias - prefers one sequence over another

Sequencing

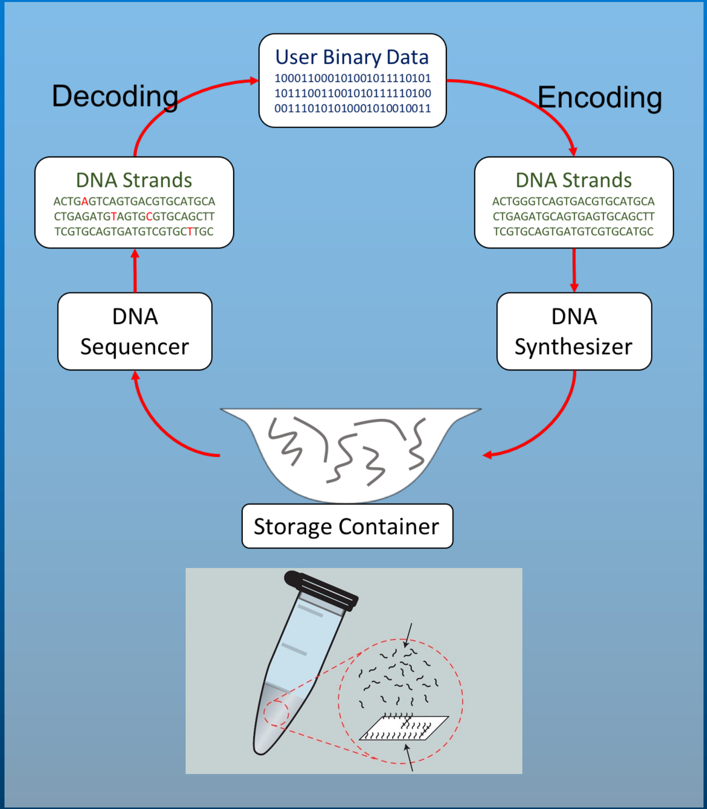
Higher GC content affects sequencing error

Presence of Homopolymers increases the error rate



TECHNOLOGY LIMITATIONS

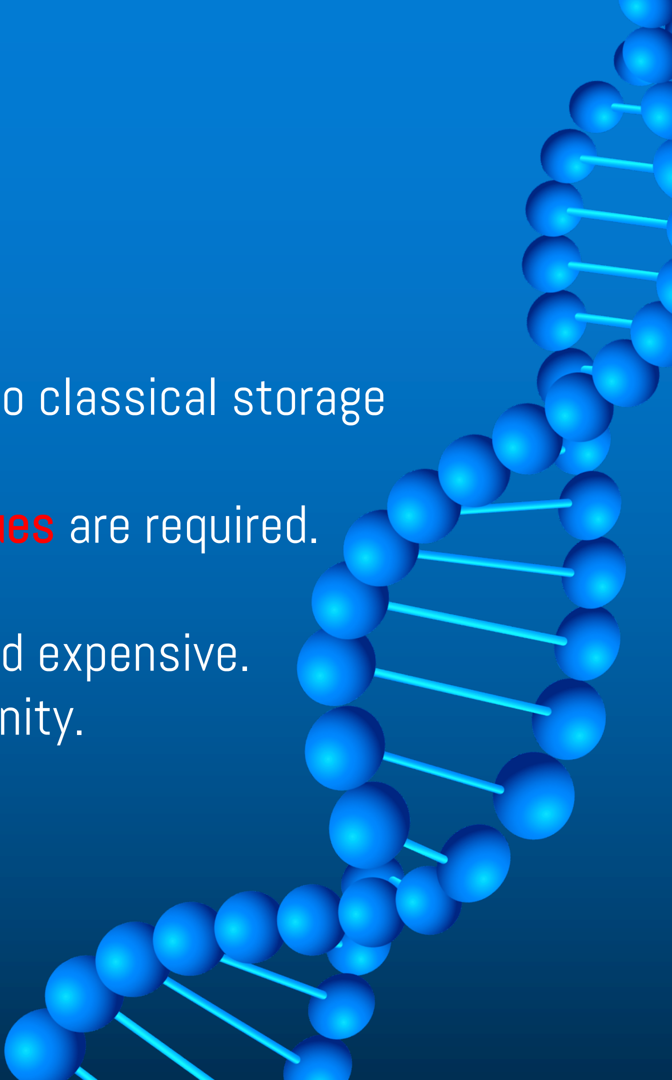
- Synthesized strands are limited in their length (roughly up to 300 symbols).
- Multiple (thousand to millions) noisy copies per designed strands.
- The noisy copies are mixed and stored together.



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.



RELATED WORK

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.

Michael Schwarz, Marius Welzel, Tolganay Kabdullayeva, Anke Becker, Bernd Freisleben, Dominik Heider, MESA: automated assessment of synthetic DNA fragments and simulation of DNA synthesis, storage, sequencing and PCR errors, Bioinformatics, Volume 36, Issue 11, June 2020, Pages 3322–3326, https://doi.org/10.1093/bioinformatics/btaa140

RELATED WORK

DEEPSIMULATOR: A DEEP SIMULATOR FOR DNA SEQUENCING

- Simulates nanopore sequencing: including the raw signal.
- Deep-learning-based tool.

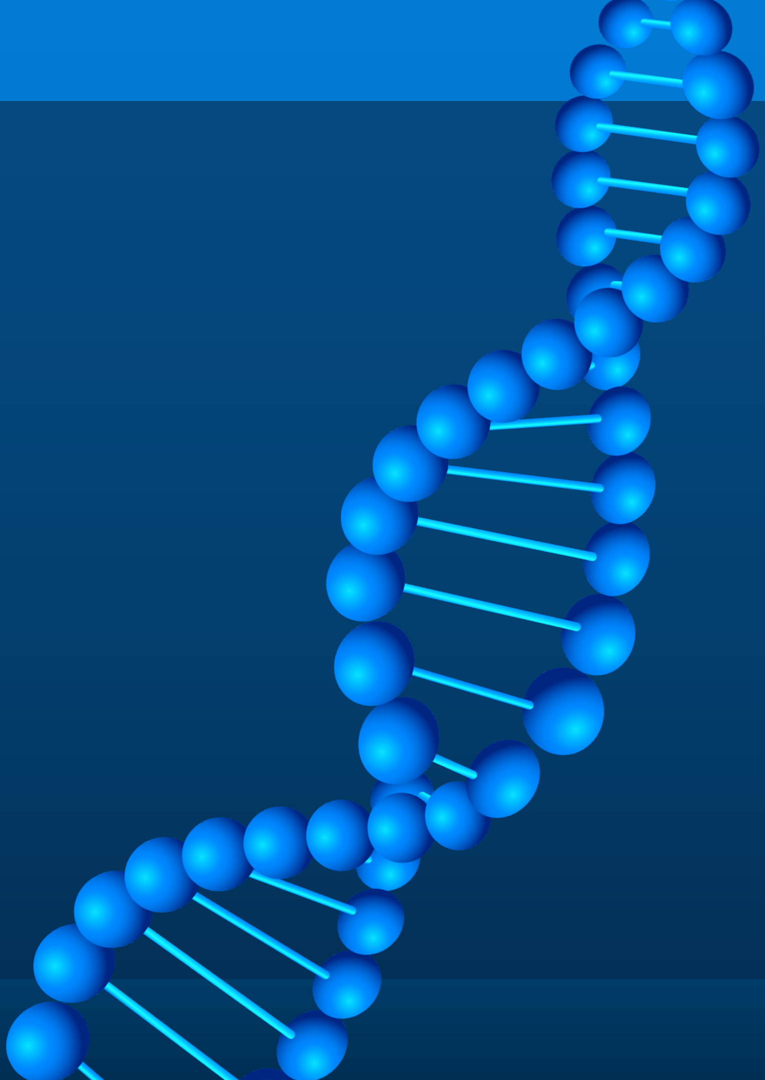
Yu Li, Renmin Han, Chongwei Bi, Mo Li, Sheng Wang, Xin Gao, DeepSimulator: a deep simulator for Nanopore sequencing, *Bioinformatics*, Volume 34, Issue 17, 01 September 2018, Pages 2899–2908, https://doi.org/10.1093/bioinformatics/bty223

RELATED WORK

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.

E. G. S. Antonio, T. Heinis, L. Carteron, M. Dimopoulou and M. Antonini, "Nanopore Sequencing Simulator for DNA Data Storage," *2021 International Conference on Visual Communications and Image Processing (VCIP)*, 2021, pp. 1-5, doi: 10.1109/VCIP53242.2021.9675388.



THE STORALATOR

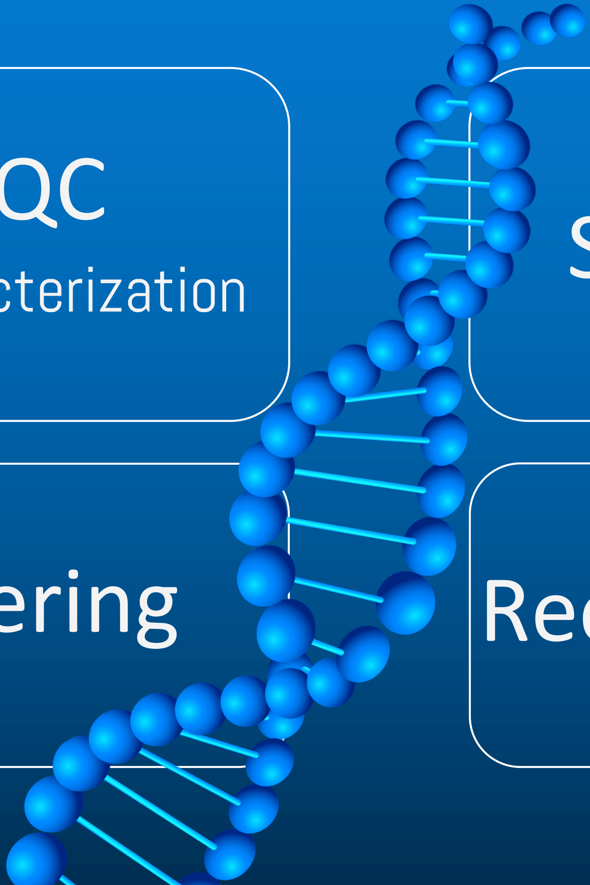
A software tool that allows researchers from all fields to compare, study, and improve their coding techniques and algorithms with current state-of-the-art solutions.

THE DNA-STORALATOR

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

Error Simulation

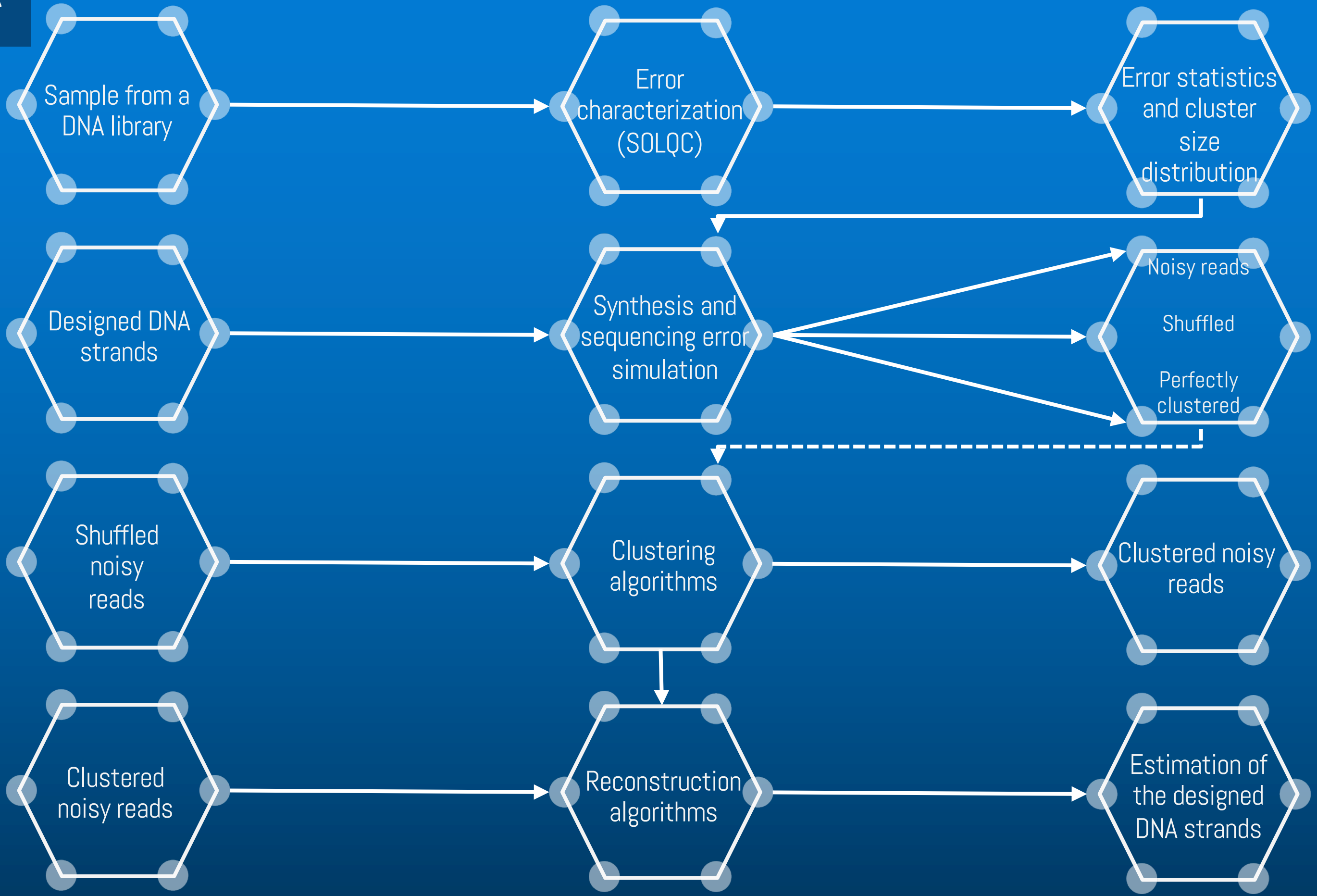
Synthesis, PCR, Sequencing

Clustering

Reconstruction

THE STORALATOR

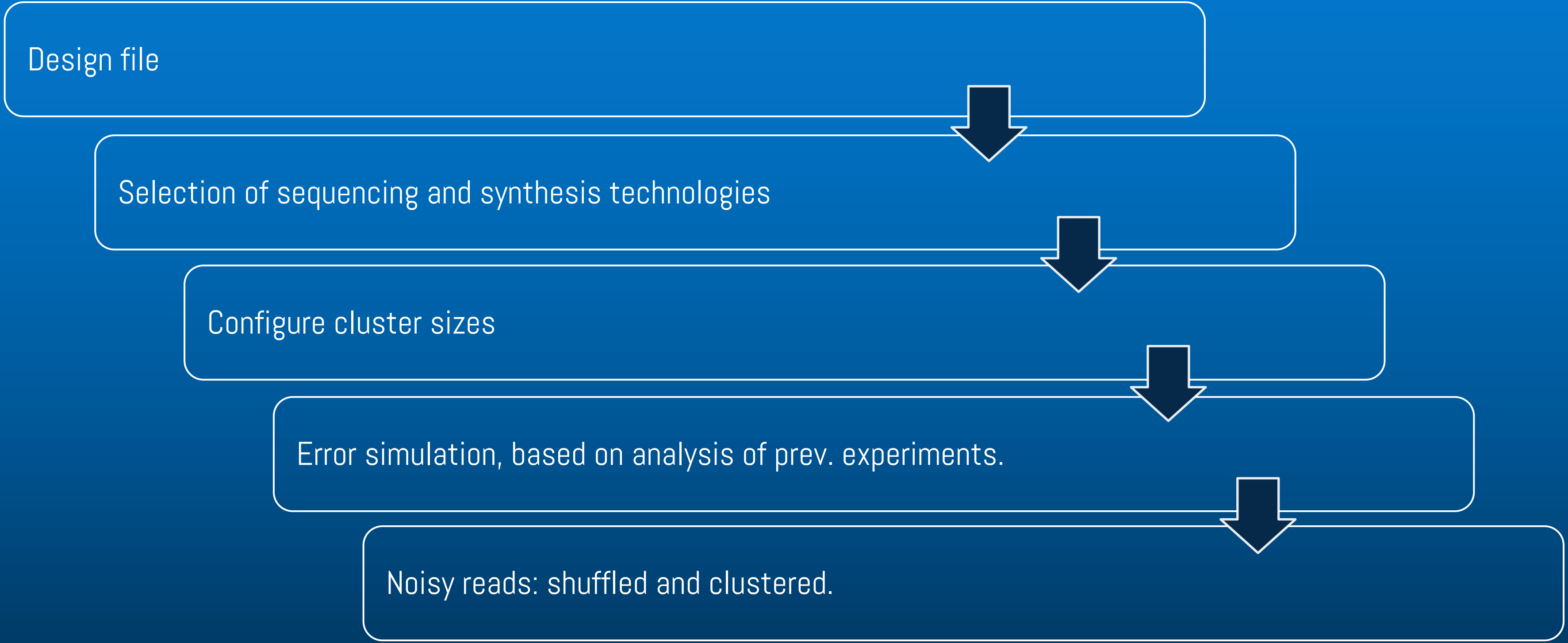
THE STORALATOR FLOWS



Sample from a DNA library
Error characterization (SOLQC)
Error statistics and cluster size distribution
Designed DNA strands
Synthesis and sequencing error simulation
Noisy reads
Shuffled
Perfectly clustered
Shuffled noisy reads
Clustering algorithms
Clustered noisy reads
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.

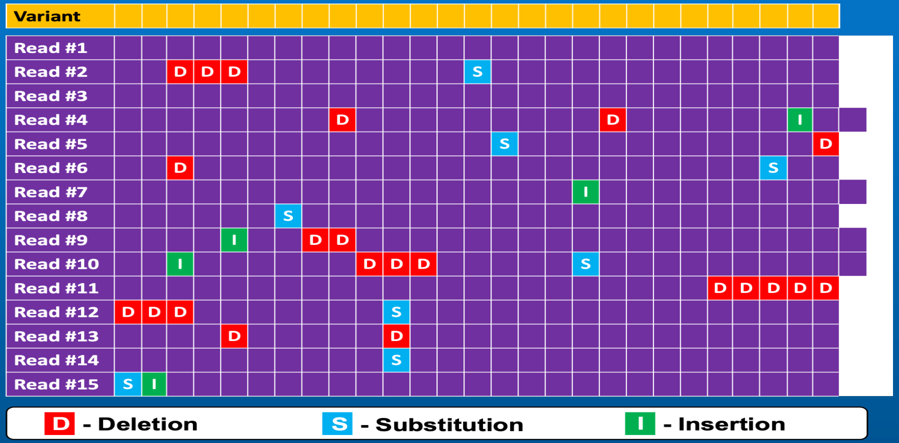
OUTPUT

SOLQC: Synthetic Oligo Library Quality Control Tool

Omer Sabary, Yoav Orlev, Roy Shafir, Leon Anavy, Eitan Yaakobi, Zohar Yakhini, SOLQC: Synthetic Oligo Library Quality Control tool, Bioinformatics, Volume 37, Issue 5, 1 March 2021, Pages 720–722, https://doi.org/10.1093/bioinformatics/btaa740

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.



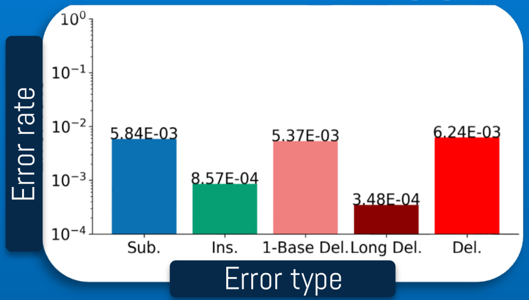
Variant																											
Read #1																											
Read #2			D	D	D									S													
Read #3																											
Read #4									D										D							I	
Read #5															S												D
Read #6			D																						S		
Read #7																		I									
Read #8							S																				
Read #9					I			D	D																		
Read #10			I							D	D	D						S									
Read #11																							D	D	D	D	D
Read #12	D	D	D								S																
Read #13					D						D																
Read #14											S																
Read #15	S	I																									

D - Deletion　S - Substitution　I - Insertion

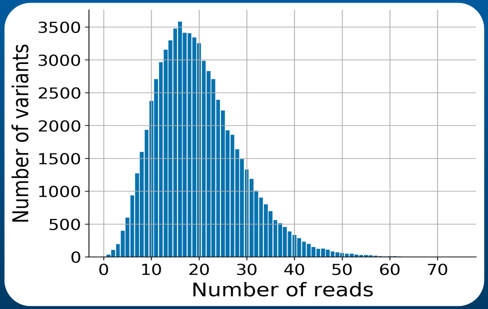
THE STORALATOR

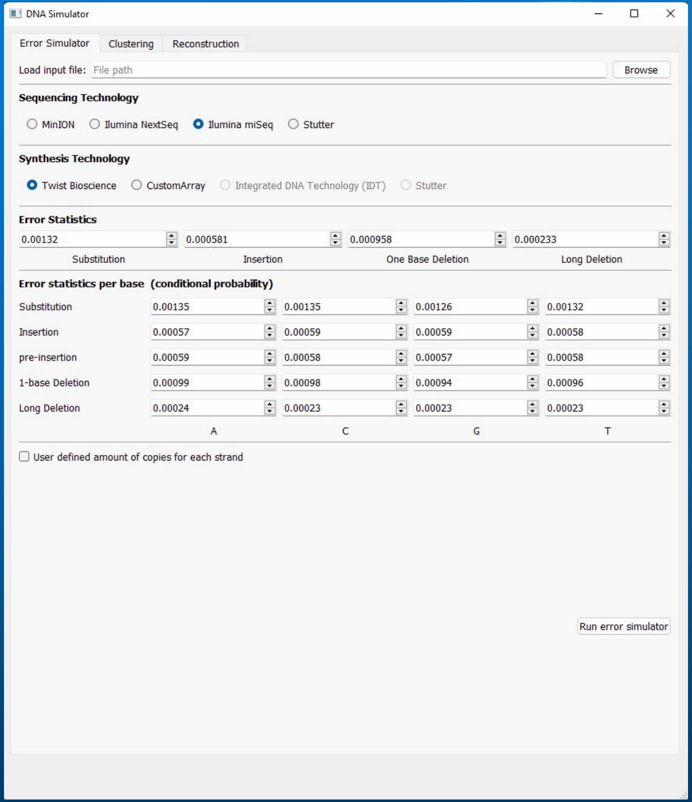
SOLQC RESULTS - EXAMPLE

ERROR RATES IN [1]



CLUSTER SIZE DISTRIBUTION [2]





Load input file: File path

Sequencing Technology

○ MinION ○ Ilumina NextSeq ● Ilumina miSeq ○ Stutter

Synthesis Technology

● Twist Bioscience ○ CustomArray ○ Integrated DNA Technology (IDT) ○ Stutter

Error Statistics

Substitution	Insertion	One Base Deletion	Long Deletion
0.00132	0.000581	0.000958	0.000233

Error statistics per base (conditional probability)

	A	C	G	T
Substitution	0.00135	0.00135	0.00126	0.00132
Insertion	0.00057	0.00059	0.00059	0.00058
pre-insertion	0.00059	0.00058	0.00057	0.00058
1-base Deletion	0.00099	0.00098	0.00094	0.00096
Long Deletion	0.00024	0.00023	0.00023	0.00023

□ User defined amount of copies for each strand

[1] Grass, Heckel, Puddu, Paunescu, and Stark, **Robust chemical preservation of digital information on DNA in silica with error-correcting codes**. Angewandte Chemie International Edition, 2015.

[2] Erlich and Zielinski, **DNA fountain enables a robust and efficient storage architecture**. Science, 2017.

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.

CLUSTERING ALGORITHMS

- 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.

[1] C. Rashtchian, K. Makarychev, M. Racz, S. Ang, D. Jevdjic, S. Yekhanin, L. Ceze, and K. Strauss
Clustering billions of reads for DNA data storage, dvances in Neural Information Processing Systems, 30.. 2017.

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

Omer Sabary, Alexander Yucovich, Guy Shapira, Eitan Yaakobi. “Reconstruction Algorithms for DNA-Storage Systems,”
bioRxiv 2020.09.16.300186; doi: https://doi.org/10.1101/2020.09.16.300186

USE CASE EXAMPLES

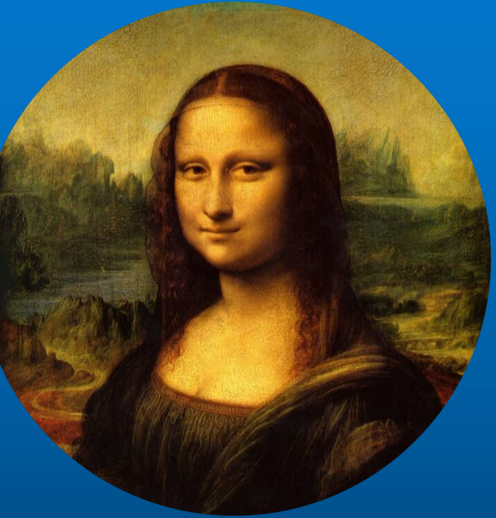
That can benefit the DNA storage community.

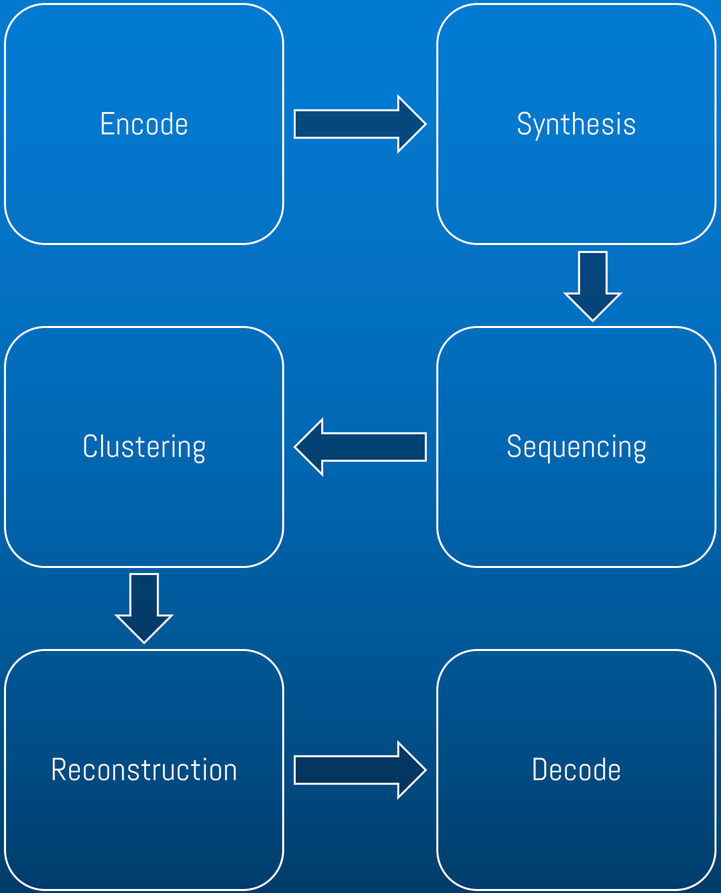
A DEVELOPMENT OF NEW CODING TECHNIQUES

- 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

Sequencing

Clustering

Reconstruction

Decode

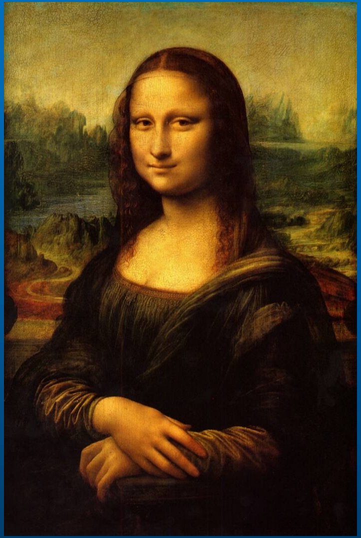
ENCODING/DECODING BY

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

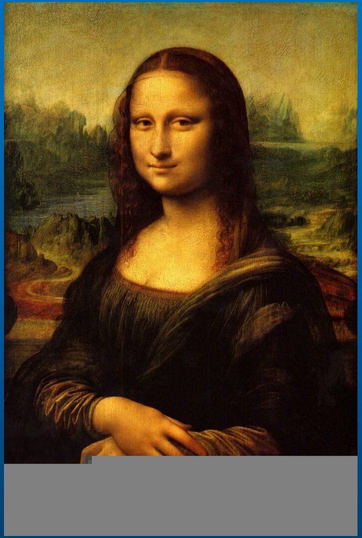
Zhi Ping, Haoling Zhang, Shihong Chen, Qianlong Zhuang, Sha Joe Zhu, Yue She

RESULTS – BLAWAT ET AL.

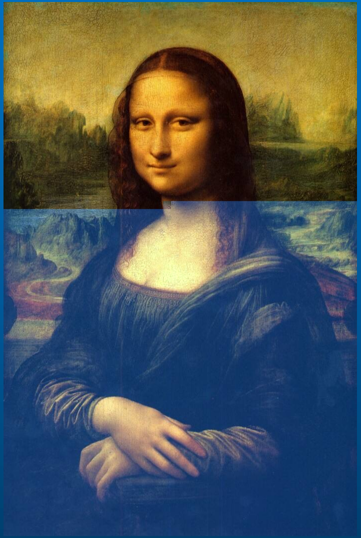
Hamming + RS



RS



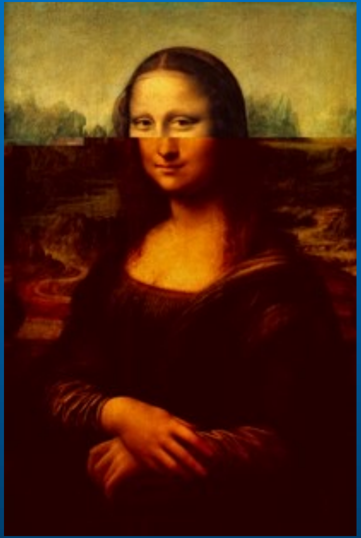
No ECC



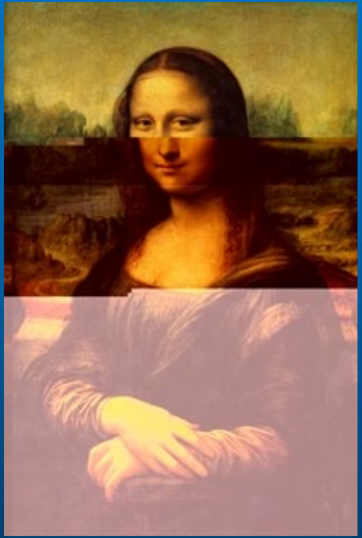
Blawat, M., Gaedke, K., Huetter, I., Chen, X. M., Turczyk, B., Inverso, S., ... & Church, G. M. (2016). Forward error correction for DNA data storage. *Procedia Computer Science*, *80*, 1011-1022.

RESULTS – GOLDMAN ET AL.

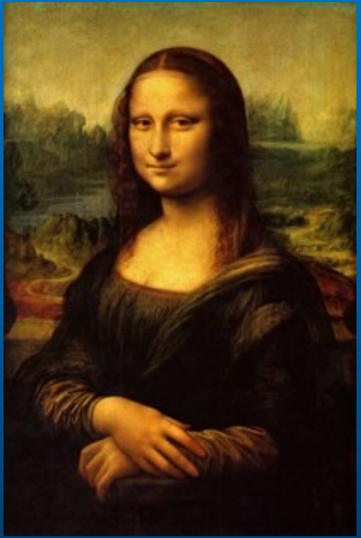
1 deletion
2 insertions
1 substitution



1 deletion



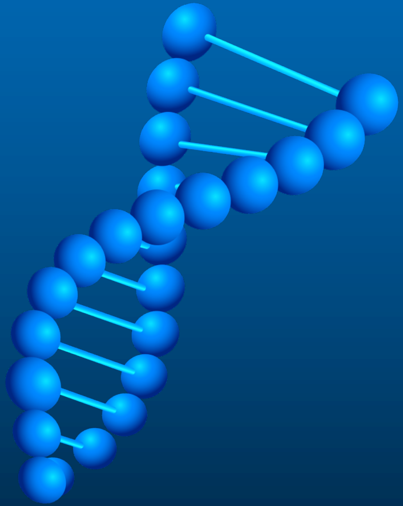
1 substitution



Goldman, Bertone, Chen, Dessimoz, LeProust, Sipos, and Birney, Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature, 2013.

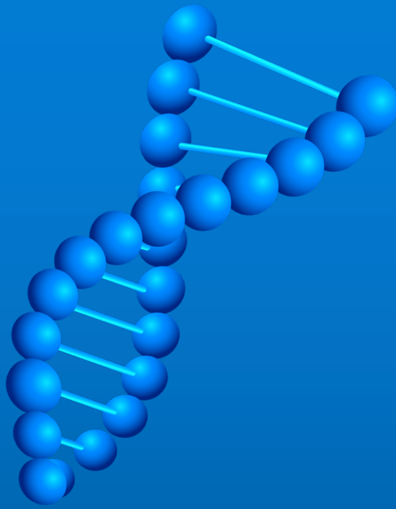
USE CASE EXAMPLES

DEVELOPMENT OF NEW ALGORITHMS FOR DNA STORAGE



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

USE CASE EXAMPLES



EXPERIMENT DESIGNING

The storalator provides an efficient method to test new algorithms and coding techniques before performing expensive and time-consuming wet experiments.

FUTURE WORK

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.

FUTURE WORK

DO YOU HAVE MORE RESULTS?

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

Thanks

SPECIAL THANKS TO

Gadi Chaykin
Nili Furman
Eitan Yaakobi
Dvir Ben-Shabat

For helping grow this project ☺

THANKS

Does anyone have any questions?