DNA Data Storage and Near-Molecule Processing for the Yottabyte Era

Karin Strauss, Microsoft Research
Luis Ceze, University of Washington
Storage capacity is growing too slowly

Source: IDC
When Moore met Feynman

The number of transistors doubles every 18-24 months

The industry roadmaps are based on that continued rate of improvement

Arrange the atoms the way we want

DNA molecules use approximately 50 atoms for one bit

Let’s store data in DNA!

DNA data storage basics

Store data in synthetic DNA strands

Simple mapping:

<table>
<thead>
<tr>
<th>Bits</th>
<th>Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>A</td>
</tr>
<tr>
<td>01</td>
<td>C</td>
</tr>
<tr>
<td>10</td>
<td>G</td>
</tr>
<tr>
<td>11</td>
<td>T</td>
</tr>
</tbody>
</table>

10000111001001

150 to 300 bases
Dense, really dense

Cold Storage: 1EB
Size: Two Walmart Supercenters
Information durability

DNA “synthetic fossils” last 2,000 – 2,000,000 years

Extreme density makes these conditions cheap and easy to keep
Comparison with other media

Potential of DNA: >10^7 improvement

Gb/mm^3

- Disk (Rot)
- Disk (SSD)
- Flash (Chip)
- Tape
- Optical
- DNA

Lifetime (years)
- 3-5
- 5
- 5
- 10-30
- 10-unlimited
- 1000+

metadata spacing redundacy
No obsolescence

Size of a mainframe
800 Kbases/day

Size of a workstation
80 Gbases/day

Size of a portable SSD
10 Gbases/day

Same medium as read technology improves:

G A C T A G C A C T
No obsolescence issue, DNA will always be relevant

Same medium as read technology improves:

Medium changes as read and write technology improves:
Polymerase Chain Reactions (PCR) create copies exponentially.
Sustainability

DNA promises to be significantly more sustainable than tape

Nguyen et al., Electronics Goes Green, 2020
DNA storage end-to-end system

Electronic

Encoding

G

C

A

C

T

(write)

Synthesis

Molecular

Random Access

Preservation

( read)

Decoding

Electronic

G

C

A

T

Sequencing

G

C

A

T

Electronic

1

0

0

1

0

1

0

1

1
Our results so far

1GB of data stored and fully recovered

Most data in DNA in peer reviewed publication

Organick et al., Nature Biotechnology, 2018
DNA storage end-to-end system

Electronic → Encoding → GC ACT → Synthesis (write) → Molecular → Random Access → Preservation → Sequencing (read) → Electronic → Decoding → GC ACT → Electronic
DNA encoding

Grass et al., Angewandte Chemie, 2015; Bornholt et al., ASPLOS, 2016
DNA synthesis

Addition

Oxidation

Deblocking

ACGT  AC  GACTCTGA  TGCA

Array synthesis
DNA preservation

(Credit: Grass et al./ETH Zurich)

Grass et al., Angewandte Chemie, 2015; Chen et al., Advanced Functional Materials, 2019; Kohll et al. Chemical Communications, 2020
Selecting one item out of two

PCR primers are used to access units of storage individually

Yazdi et al., Scientific Reports, 2015; Bornholt et al., ASPLOS, 2016; Organick et al., Nature Biotechnology, 2018; Organick et al., Nature Communications, 2020; Chen et al., Nature Communications, 2020
Reading DNA with sequencing by synthesis
Reading DNA with nanopores

Yazdi et al., Scientific Reports, 2017; Organick et al., Nature Biotechnology, 2018; Lopez et al., Nature Communications, 2019
Different error profile across platforms

Illumina NextSeq

ONT MinION
DNA decoding and error correction

Rashtchian et al., NeurIPS, 2017; Organick et al., Nature Biotechnology, 2018
DNA decoding and error correction

```
ACGT  AC  GACTCTGA  TGCA

0001 0101000101011100

ACGT  GA  GCTACAGC  TGCA

1100  1001010001001001

0100  1001010001001001

0010  1010000101011100

0011  1101100001111000

0011  0111010111001000

1100  1001010001001001

0100  1001010001001001

0010  1101100001111000

0010  1101100001111000

0011  0111010111001000

0011  0111010111001000
```
DNA decoding and error correction
DNA storage end-to-end system
Exponential improvements in DNA data storage

[Ceze et al., Nature Genetics Reviews'19]
Performance of reading and writing DNA

Carlson’s Curves - Throughput

Latency

Synthesis and sequencing are currently batch processes, matches archival storage SLAs (~hours).

Emerging technologies, like nanopore devices, provide closer to real time latency.

Source: Robert Carlson
Write and read mechanisms
Opportunities for improvements in write and read throughput, latency and cost

<table>
<thead>
<tr>
<th>Opportunities</th>
<th>Life sciences</th>
<th>Data storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate</td>
<td>Single base mutations affect function</td>
<td>Error correcting codes allow data recovery even in the presence of multiple errors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error types: substitutions, deletions and insertions</td>
</tr>
<tr>
<td>Length (“block size”)</td>
<td>Longer sequences have more function</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shorter sequences are faster and easier to make</td>
</tr>
</tbody>
</table>

![Graph showing DNA strand length vs. Overhead]

- **Optimizations**
  - **Error rate**: Single base mutations affect function. Error correcting codes allow data recovery even in the presence of multiple errors.
  - **Length (“block size”)**: Longer sequences have more function. Shorter sequences are faster and easier to make.

**Diminishing returns**

**Sweet spot**
DNA storage end-to-end system w/ integrated computing

Highly parallel and energy efficient
DNA computing in the 80s

**Problem:** shifts complexity from time to amount of material
DNA "computing" in the age of big data

Operate over data already stored in DNA
Target polynomial time algorithms
Extremely parallel and energy efficient
Content-based image/video search
Content-based image/video search in DNA
Exploiting matches for exact and approximate search

Double helix: complete match

Good partial match

Poor partial match
Searching with DNA

Match-dependent yield

3x Perfect match

2x Good partial match

1x Poor partial match
Content based media search

Database/ training

- image 2
- image 5
- image 6
- image 8

Query/ inference

Stewart et al., DNA24, 2018; Nature Communications, 2021
"Database" strands  
(many)

ACGATTTGCAATTGCA

ACGTAGACCCACCTTTAA

TCTGACATCAGACATCTCT

TGGCCACCATGACCTCCT

ATGCCACCTTTGCAATTGCA

CTGCCCATACTACGACC

"Query" strands  
(one type)

ACGATTTGCAATTGCA

ACGTAGACCCACCTTTAA

TCTGACATCAGACATCTCT

TGGCCACCATGACCTCCT

ATGCCACCTTTGCAATTGCA

CTGCCCATACTACGACC
“Semantic” Hashing

How to do this encoding?

Similar in Feature Space (Euclidean)

Images

Binary Addresses

Similar in Address Space (Hamming)

DNA Sequences

High Extraction/Reaction Yield

Adapted from Salakhutdinov et al. 2007
### Learning-based encoding

<table>
<thead>
<tr>
<th>Layer</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence Output</td>
<td>30 x 1</td>
</tr>
<tr>
<td>ReLU + Softmax Activations</td>
<td>30 x 4</td>
</tr>
<tr>
<td>Fully Connected Weights</td>
<td>10 x 128 x 30 x 4</td>
</tr>
<tr>
<td>ReLU Activations</td>
<td>10 x 128</td>
</tr>
<tr>
<td>Convolutional Weights 2</td>
<td>1 x 128 x 128</td>
</tr>
<tr>
<td>Sine Activations</td>
<td>10 x 128</td>
</tr>
<tr>
<td>Convolutional Weights 1</td>
<td>1 x 128</td>
</tr>
<tr>
<td>Input Features</td>
<td>10 x 1</td>
</tr>
</tbody>
</table>
Experiments show encouraging results

1.6M images, 3 queries

Magnetic bead extraction

Illumina-based sequencing
Physically “diffusing” computation through data offers parallelism and virtually unlimited access bandwidth. Yes, at a higher latency.

Carmean et al., IEEE Proceedings, 2018

Yottabyte-scale near-molecule computing?

Capacity/bandwidth going up

Near-data processing

Capacity \sim L^3

Bandwidth \sim L^2

Physically “diffusing” computation through data offers parallelism and virtually unlimited access bandwidth. Yes, at a higher latency.
DNA storage end-to-end system w/ integrated computing

Highly parallel and energy efficient
End-to-end system in a datacenter
First fully automated DNA data storage system

Takahashi et al., Nature Scientific Reports, 2019
First fully automated DNA data storage system

Takahashi et al., Nature Scientific Reports, 2019
Digital microfluidics

Versatile platform to implement wet lab preparation protocols

Newman et al., Nature Communications, 2019; Willsey, ASPLOS, 2019; Stephenson et al., IEEE Micro, 2020
Random access with spots + digital microfluidics

60s dwell time
33ng mass

No measurable contamination

[Stephenson, Takahashi, Nguyen, et al., Nature Communications'19]
def thermocycle(droplet, temps_and_times):
    for temp, time in temps_and_times:
        heat(droplet, temp, time)
    if droplet.volume < MIN_VOLUME:
        droplet += input("water", min_volume)

def pcr(droplet, n_iter):
    thermocycle(droplet, n_iter * [
        (95, 3 * minutes),
        (62, 30 * seconds),
        (72, 20 * seconds),
    ])
Hardware, software, wetware

Molecular domain

Electronic domain
Future hybrid systems

Biomolecules: Self assembly, massive data and efficiency

Electronics: Ultra low latency, engineerable, perfect control

Quantum: Massive specialized parallel computing, little data
Questions?

https://misl.cs.washington.edu
https://www.microsoft.com/en-us/research/project/dna-storage/