DNA data storage and computation

Karin Strauss, Microsoft
Luis Ceze, University of Washington
DNA data storage advantages

Density

Durability

No obsolescence issue

Ability to perform computation
DNA data storage basics

Bases: A C G T

Data: 10000111001001

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>

Store data in synthetic DNA strands

150 to 300 bases
Improvements in DNA data storage

Ceze, Nivala and Strauss, Nature Reviews Genetics, 2019
DNA storage end-to-end system
End-to-end system in a datacenter
First fully automated DNA data storage system

synthesis

sequencing

digital microfluidics

storage & prep

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
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),
    ])

High-level programming with Puddle

"Assembly code"

Hardware

Willsey et al., ASPLOS, 2019
DNA storage end-to-end system
Reading DNA with nanopores

Yazdi et al., Scientific Reports, 2017; Organick et al., Nature Biotechnology, 2018; Lopez et al., Nature Communications, 2019
Reading DNA with nanopores

Yazdi et al., Scientific Reports, 2017; Organick et al., Nature Biotechnology, 2018; Lopez et al., Nature Communications, 2019
DNA computing in the 80s

Hamiltonian path problem

Problem: shifts complexity from time to amount of material

Adleman, DNA1, 1994
DNA "computing" in the age of big data

Operate over data already stored in DNA
Target polynomial time algorithms
Extremely parallel and energy efficient
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

AGTCAGCGTGACTACGTAG
ACGTA

image 5

CTGAGTAAAGTCGCGTAGTA
TCGAT

image 6

GCGTACTTAGCTGATGACT
GTCAG

image 8

TGCACGTACTCGTAGCTGA
AGTCA

Query/ inference

CTGTGTAAGTCGTAATAGTA

GACACATTCCAGCATATCAT

Stewart et al., DNA24, 2018
Content based media search

Database/ training

image 2

image 5

image 6

image 8

Query/ inference

CTGAGTAAGTCGCGTAGTA
GACACATTCAGCATATCAT

Stewart et al., DNA24, 2018