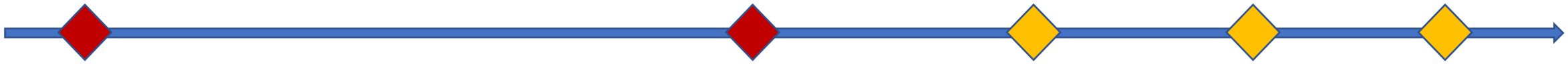
Coding for efficient DNA data storage

Sergey Yekhanin

Microsoft

Based on joint work with: Luis Ceze, Parikshit Gopalan, Sivakanth Gopi, Konstantin Makarychev, Henry Pfister, Miklos Racz, Amir Shahrasbi, Sundara Srinivasvaradhan, Cyrus Rashtchian, Karin Strauss

DNA data storage



1965:
Mikhial Neiman:
The idea of storing data in synthetic DNA

2010:
George Church et al.:
The first DNA data storage system (3Kb)

2016:
Erlich and Zielinksi:
2.11MB

2017:
Blawat et al.:
22MB

2018:
Organick et al.:
200MB

- Synthetic DNA provides unparalleled density and storage lifespan: cold archival storage
- Key issues: cost of read / write operations and throughput of DNA data storage systems

- Very active area of research in academia and industry

illumina® TWIST BIOSCIENCE Microsoft WD Western Digital® CATALOG INFINITE DATA ARCHIVES ANSA BIOTECHNOLOGIES

Outline

- DNA data storage channel
- High level view of encoder and decoder
- Key metrics (cost of read and write)

DNA Storage: the channel

Writing:

Encoder

A set of short strings over {A,C,G,T}

DNA synthesis

A pool of DNA strands

Reading:

DNA sequencing

Multiple noisy reads of every stored DNA strand

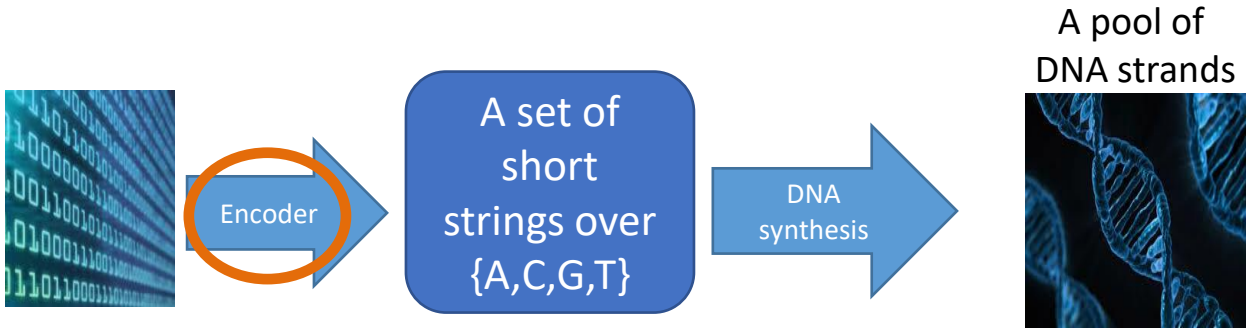
Decoder

Errors:

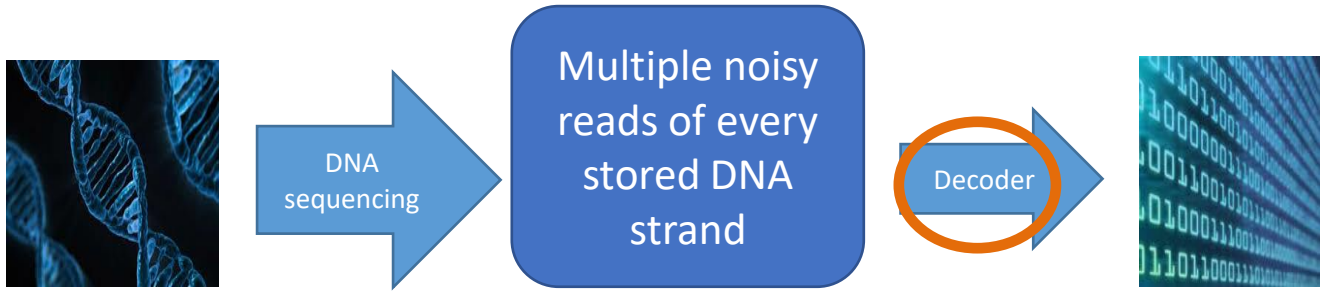
- Some strands fail to be synthesized or get zero reads in sequencing (erasures).
- Strands may have insertions / deletions / substitutions of bases.
- Error rate of $1\% - 10\%$ per coordinate depending on sequencing technology (Illumina / Nanopore).
- Observing multiple reads of a strand.

DNA Storage: the channel

Writing:



Reading:



Parameters:

- Strand length of 100 – 200
- Pool size: millions -> billions

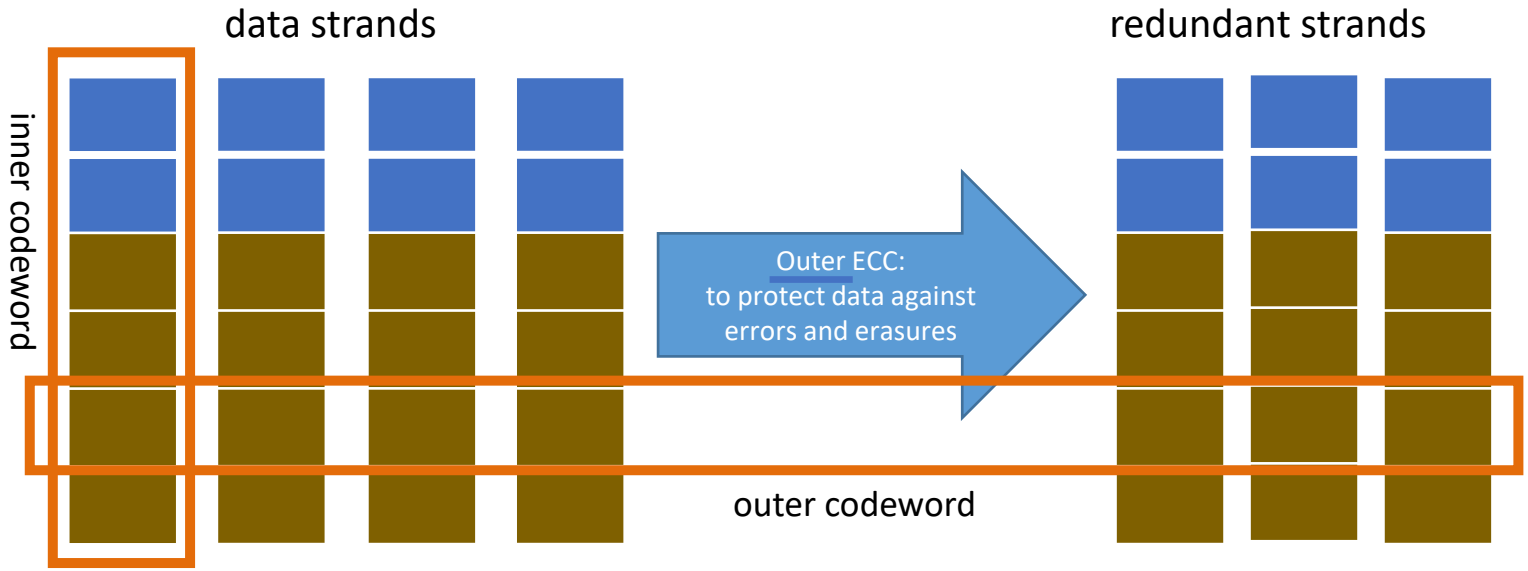
Key metrics:

- Cost of read: bits-per-base on read
- Cost of write: bits-per-base on write (**rough measure**)
- Fast encoding / decoding

Generic codec architecture: Encoder

1. Randomize data using a pseudorandom number generator / encrypt / compression
2. Represent data in strands:

3. Generate redundant strands using a standard ECC (Reed-Solomon, LDPC):



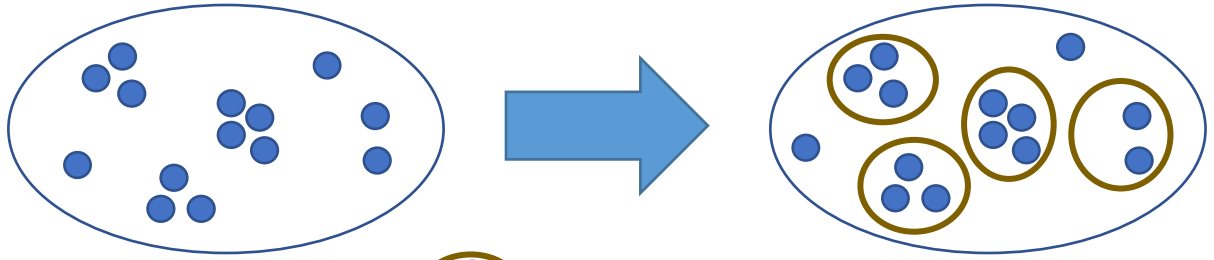
4. Possibly:

   Use inner code to additionally guard individual strands (to reduce Hamming errors).

   Ensure constrained representation (long runs of repeated characters, GC content, DNA secondary structure). Details.

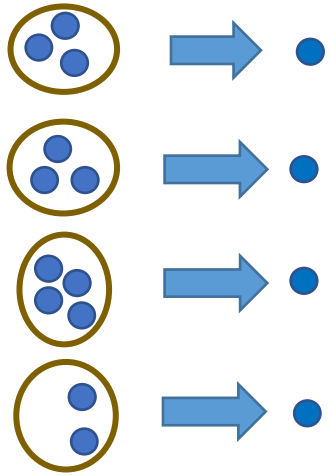
Codec architecture: Decoder

1. Clustering:



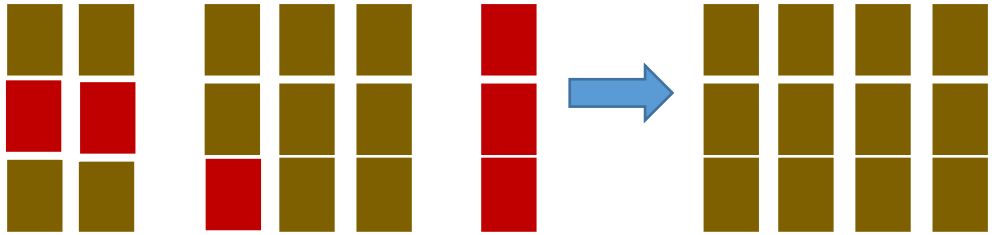
- No clustering
- Based on prefix
- Based on full strand content [RMR+17]

2. Trace reconstruction / inner decoding:



- No inner codes
- Ad-hoc codes, RS codes
- Convolutional codes for edit distance channel [SDDF'19, LMW+'20, GSPY'21]

3. Decoding the outer code:



Data and redundant strands with errors and erasures

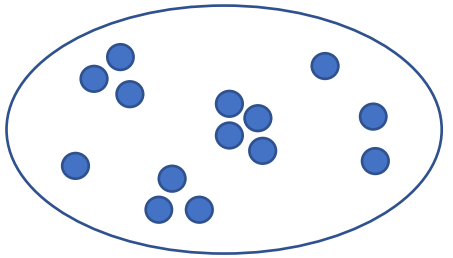
Corrected data strands

4. Remove the randomization

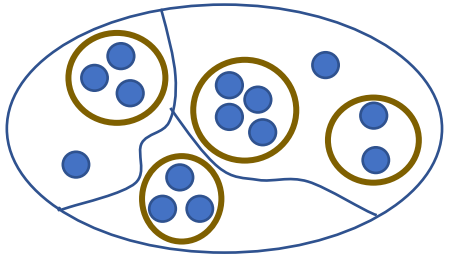
Clustering

- **Problem setup:** Edit distance, billions of points, data is well clustered into tiny clusters.

- **Trivial algorithm**:
  - Compute distances between all pairs of strands.
  - Identify strands that are close as belonging to the same cluster.



- **Our algorithm [RMR+'17]**:
  Initially: Treat every strand as a separate cluster.
  Iteratively:
    - Pick a hash function f that maps strands to buckets.
    - Partition strands according to the value of f.
    - In each bucket:
      - Compute all pairwise distances,
      - Identify close strands as belonging to the same cluster.



- **Key ingredient**: a carefully designed family of hash functions $\{f\}$, where functions $\{f\}$ tend to map close strands to the same bucket.

Cost of DNA synthesis

DNA synthesis: the main bottleneck in the DNA storage pipeline: High cost / Low throughput

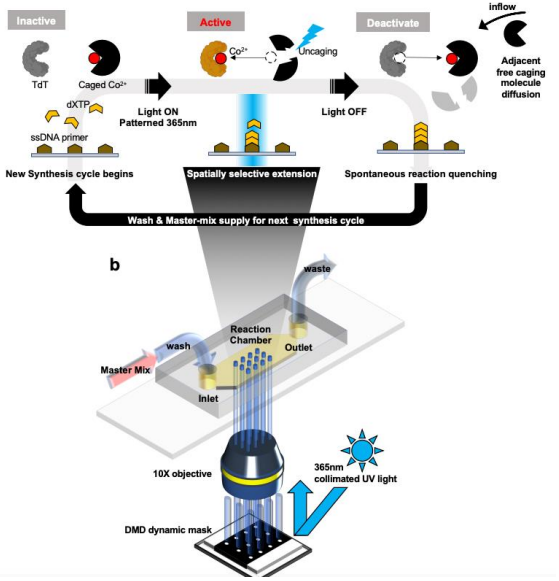
- Current technology: array-based synthesis
- Setting via an example:

Reference strand R:

	A	G	C	T	A	G	T
$S_1 = AGCT$	A	G	C	T	-	-	-
$S_2 = GCAT$	-	G	C	-	A	-	T
$S_3 = CTAG$	-	-	C	T	A	G	-

$cost_R(S_1) = 4 \qquad cost_R(S_2) = 7 \qquad cost_R(S_3) = 6$

$\mathcal{S} = \{S_1, S_2, S_3\}$ $\quad cost_R(\mathcal{S}) = \max\{cost_R(S_1), \dots, cost_R(S_3)\} = 7$



[Lee et al., 2020]

- Pools of DNA strands that share a short common super-sequence are cheap to synthesize
- Batch optimization has significant potential for reducing the cost [MRRY'20]

Conclusions

- DNA data storage: an emerging storage technology
- Synthesis / sequencing / coding for DNA data storage are all advancing rapidly

- To early to fix the encoding format. May stifle innovation.
- One possibility: fix the language to specify the encoding format
  (like with SGML: Standard Generalized Markup Language)