Part 1: DNA-Based Data Storage and Computing

Olgica Milenkovic
University of Illinois, Urbana-Champaign

North American School of Information Theory, Texas, 2018

May 2018
The Era of Massive Data

- Large Hadron Collider: 600 million collisions/s, 0.5 PB per week.
- DNA sequencing data: 30 – 50 TB per week.
- Sloan Digital Sky Survey: 1 – 2 TB per week.
- Social networks (Twitter, Facebook, LinkedIn), NASA weather surveys, consumer and stock market data, Internet sources...

Credit: In search of the God particle, Wikipedia.
How to Cope?

- **Pushing the limits of existing storage media:** Magnetic tapes, disks, flash, 3D flash, ...
- **Data compression:** New initiatives by NIH (BD2K Targeted Software Development for Genomic Data Compression) and other efforts.
How to Cope?

- **New storage media**: quantum memories, nanofilm storage, polymer-based storage?
- **Data compression**: What densities are possible?
DNA-Based Data Storage
Looking for Alternative Storage Media: DNA

- **DNA is extremely durable:** Can still “read” mammoth, Neanderthal, and 700,000 old horse bone DNA!
- **DNA information content of Human cell:** 6.4 GB. **Mass of a cell:** ~3 picograms. **No. of cells:** 15 – 40 × 10^{12}.
- **Can one store information in DNA?**
- **This question has been raised before:** “There is plenty of room at the bottom,” R. Feynman.
We can write…

We can “write” in DNA using what is called the process of DNA Synthesis.

**Biochemistry of synthesis:** Stitching together bases from the set \( \{A, T, G, C\} \) through deprotection & coupling cycles.
**Commercial synthesis:** Agilent, Gen9, IDT, Twist Bioscience (recent feature on “Rewriting DNA Synthesis on Silicon”).

- **DNA microarray-based short string (oligo) pool synthesis (left):** Cost effective, large scale. Moderate error rates.
- **Long strand (gBlocks) synthesis (right):** Assembles short blocks. Chemical error-correction.
- **Types of synthesis errors:** Deletions, insertions, substitutions.
We Can Read...

**Illumina (MiSeq):** Best overall performance of modern sequencing technologies in terms of yield and accuracy. Relatively small error rates (substitutions and rare deletions). Short read length.

**Steps:** Cloning /// Shearing /// Reading of unordered pool /// Computer aided alignment of overlapping fragments /// Consensus
We Can Read...

**Oxford Nanopore - MinIon:** Longer read lengths, portable architecture. Context-dependent deletion, insertion and substitution errors.

**Key properties:** Biological pore(s) and motor, base calling using deep learning techniques.
We Can Amplify and Enable Random Access...

**Polymerase Chain Reaction (PCR):** Cheap, fast, “exponential” information replication.

**Primers - Key enablers of PCR:** Short DNA strands that initiate replication at “strand-matching” locations (red blocks).
We Can Amplify and Enable Random Access...

**Polymerase Chain Reaction (PCR):** Cheap, fast, “exponential” information replication.

**Primers - Key enablers of PCR:** Short DNA strands that initiate replication at “strand-matching” locations (red blocks).
DNA Data Storage Platforms
Church *et al.* (Science, 2012) and later Goldman *et al.* (Nature, 2013) stored 739 KB of data in synthetic DNA, mailed it and recreated the original digital files using Illumina readers.

**Digital archival storage systems** that will safely store the equivalent of one million CDs in a gram of DNA for 10,000 years.
Can one Randomly Access and Rewrite the Data?

- **Problem 1:** Random access was impossible in first implementation - need to “read” whole book to find one sentence.
- **Problem 2:** To perform editing, need to change large number of reads (fragments).
- **Problem 3:** The first schemes were sensitive to contextual errors.

Storage format of Goldman et al.: overlapping reads akin to sequencer output.

AAATTTTGCGCTATTGCCCAATTGCCGGTTAAAAATATATGAGACTCTAAAA...
A fully operational random access and rewritable DNA-based memory with *Sanger sequencing*.

- Yazdi et.al., 2015 - First random access, rewritable DNA-based storage system. Encoded Wikipedia entries for six US universities.

Random Access via Addressing and PCR

- The addressing system: Primers = Addresses, used in PCR reaction. Random access equals exponential amplification.

- How to avoid addressing errors?
Address Properties

- **Addresses need to be sufficiently different** (Hamming, Levenshtein distance) and avoided elsewhere in the blocks.
- **Addresses should not fold**: Needed for accurate amplification.
- **Addresses should have balanced GC content**: Needed for stable melting temperature.
GC Imbalance Hurts!

- Synthesis constraints identification with IDT.
- **Balancing constraints**: GC-content has to be balanced in small block-lengths at the 3' and 5' ends of the strings, longer blocks allowed within the sequence (blocklength=8).
The Constrained Coding Components

**Definition.** A sequence $a = (a_1, \ldots, a_n) \in \mathbb{F}_q^n$ is self uncorrelated if no proper prefix of $a$ matches its suffix, i.e., $(a_1, \ldots, a_i) \neq (a_{n-i+1}, \ldots, a_n)$, for all $1 \leq i < n$.

**Extension:** A mutually uncorrelated (cross-bifix-free) code is a set of sequences such that for any two sequences $a, b \in \mathbb{F}_q^n$ in the code no proper prefix of $a$ appears as a suffix of $b$ and vice versa [L70, G60, B12].
Address Sequence Construction

Enumeration and construction of strings of a given length that contain no elements of some fixed set of strings as subwords [GO80’s]:

Take addresses as “forbidden words” to ensure specific random access. Relax constraints.

MU vs. Weakly MU: A \( k \)-weakly mutually uncorrelated (WMU) code is a set of sequences such that for any two sequences \( a, b \in \mathbb{F}_q^n \) in the code no proper prefix of \( a \) of length \( \geq k \) appears as a suffix of \( b \) and vice versa [TKM16].

Construction of balanced WMU codes Hamming distance constraints?

For \( a = (a_1, \ldots, a_s), b = (b_1, \ldots, b_s) \in \{0, 1\}^s \), define

\[
\Psi(a, b) : \{0, 1\}^s \times \{0, 1\}^s \to \{A, T, C, G\}^s
\]

according to:

\[
\text{for } 1 \leq i \leq s, \; c_i = \begin{cases} 
A & \text{if } (a_i, b_i) = (0, 0) \\
C & \text{if } (a_i, b_i) = (0, 1) \\
T & \text{if } (a_i, b_i) = (1, 0) \\
G & \text{if } (a_i, b_i) = (1, 1) 
\end{cases}
\]
Decoupling the construction: Let $C_1, C_2 \subseteq \{0, 1\}^s$ be two binary block code of length $s$. Encode all pairs $(a, b) \in C_1 \times C_2$ using $C_3 = \{ \Psi(a, b) \mid a \in C_1, b \in C_2 \}$. Then:

1. $C_3$ is balanced if $C_2$ is balanced.
2. $C_3$ is a $k$-WMU code if either $C_1$ or $C_2$ is a $k$-WMU code.
3. If $d_1$ and $d_2$ are the minimum Hamming distances of $C_1$ and $C_2$, respectively, then the minimum Hamming distance of $C_3$ is at least $\min(d_1, d_2)$.

See also [LY17] for MU codes.

Information sequence encoding?
Information Sequence Encoding for Texts

Modification based on [WI90’s], and new approach in [TGM17].
Random Access and Rewriting Experiments

- Random access achieved via PCR, addresses used as primers.
- Context identification and rewriting performed via gBlock or OE-PCR methods.

![PCR of five primers and PCR of selected string from pool (A) and in individual well (B) with gBlock process diagram.](image-url)
Random Access and Rewriting Experiments

- Random access achieved via PCR, **addresses used as primers**.
- Context identification and rewriting performed via gBlock or OE-PCR methods.

Cheap 80-primer sequential rewriting

A) Two PCR products of rewrite. B) The generated PCR rewrite with correct size of 1kb.

Sequencing results of 10 plasmids (5 from original, 5 from rewrite) with primer in forward direction of the insert. The rewritten region is covered in the red square.
MinION Oxford Nanopore (R7): Sequence traces (reads).
Major Problem: Very large number of sequence-dependent indel and substitution errors (R.7 flowcell, ~ 10%, R 9.4 flowcell ~ 4%)!

Example Statistics:

<table>
<thead>
<tr>
<th>Block (length)</th>
<th>Number of reads</th>
<th>Sequencing Coverage depth</th>
<th>Number of errors: (substitution, insertion, deletion)</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Per read (average)</td>
<td>Nanopolish</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (1,000)</td>
<td>201</td>
<td>176.145</td>
<td>(107, 14, 63)</td>
<td>(14,32,5)</td>
</tr>
<tr>
<td>2 (1,000)</td>
<td>407</td>
<td>315.521</td>
<td>(123, 12, 70)</td>
<td>(75,99,40)</td>
</tr>
<tr>
<td>3 (1,000)</td>
<td>490</td>
<td>460.375</td>
<td>(80, 23, 42)</td>
<td>(10,45,0)</td>
</tr>
<tr>
<td>4 (1,000)</td>
<td>100</td>
<td>81.763</td>
<td>(69, 18, 37)</td>
<td>(1,54,1)</td>
</tr>
<tr>
<td>5 (1,000)</td>
<td>728</td>
<td>688.663</td>
<td>(88, 20, 48)</td>
<td>(4,45,3)</td>
</tr>
<tr>
<td>6 (1,000)</td>
<td>136</td>
<td>120.907</td>
<td>(79, 21, 42)</td>
<td>(390,102,61)</td>
</tr>
<tr>
<td>7 (1,000)</td>
<td>577</td>
<td>542.78</td>
<td>(83, 26, 41)</td>
<td>(3,31,3)</td>
</tr>
<tr>
<td>8 (1,000)</td>
<td>217</td>
<td>199.018</td>
<td>(83, 20, 46)</td>
<td>(18,51,1)</td>
</tr>
<tr>
<td>9 (1,000)</td>
<td>86</td>
<td>56.828</td>
<td>(60, 16, 30)</td>
<td>(404,92,54)</td>
</tr>
<tr>
<td>10 (1,000)</td>
<td>442</td>
<td>396.742</td>
<td>(91, 18, 52)</td>
<td>(388,100,59)</td>
</tr>
<tr>
<td>11 (1,000)</td>
<td>114</td>
<td>101.826</td>
<td>(79, 23, 42)</td>
<td>(16,23,18)</td>
</tr>
<tr>
<td>12 (1,000)</td>
<td>174</td>
<td>162.559</td>
<td>(94, 23, 50)</td>
<td>(14,59,1)</td>
</tr>
<tr>
<td>13 (1,060)</td>
<td>378</td>
<td>352.35</td>
<td>(88, 26, 44)</td>
<td>(7,55,4)</td>
</tr>
<tr>
<td>14 (1,000)</td>
<td>222</td>
<td>189.918</td>
<td>(69, 22, 34)</td>
<td>(15,34,3)</td>
</tr>
<tr>
<td>15 (1,000)</td>
<td>236</td>
<td>222.967</td>
<td>(92, 24, 45)</td>
<td>(15,46,2)</td>
</tr>
<tr>
<td>16 (1,000)</td>
<td>198</td>
<td>182.99</td>
<td>(103, 16, 61)</td>
<td>(15,62,4)</td>
</tr>
<tr>
<td>17 (880)</td>
<td>254</td>
<td>240.273</td>
<td>(77, 19, 42)</td>
<td>(359,95,44)</td>
</tr>
</tbody>
</table>
Sequence Alignment

- **First Step:** “Merge traces” into one consensus sequence.
Sequence Alignment

- **First Step:** “Merge traces” into one consensus sequence.

- **Bioinformatics:** Sequence alignment [NW, SW].
  **Computer Science:** Reconstructing sequences from traces [Batu et al. 2004].
**Sequence Alignment**

- **DP:** Optimal, but of very high complexity.  
  Works poorly on real data!

- **Reference-free sequence alignment:** CLUSTAL, KALIGN, MUSCLE, TCOFFEE, etc.

- **“Simple” trace reconstruction:** [Batu et.al. 2004], [Holenstein et.al. 2016]
How Do We Handle Deletions?

- **Key idea**: Use addresses (tags) as pilot sequences to identify good quality reads.
- Align good reads. Use balancing property to correct errors, repeat. Use homopolymer coding.
Encoded images in compressed format into DNA.

- Compression, Base64 conversion, error-correcting and constrained coding (balancing GC content and forbidden address sequences).
- Careful address design.

JPEG and Base64 Encoded Image

100010111001010...

Conversion into a sequence over the alphabet \{A,T,G,C\}

AATGCAGCGTTTAGAGAT...

Parsing into blocks of length 896

AATGCAGC|GTTTAGAG|AT...

Block GC-Content Balancing and ECC

AATG|CCCAGCATAG| GTTCGAGAGCTCT

Addressing

CGAA|AATG|CCCAGCATAG| GCTAGTTTCGAGAGCTCT

Data Encoding

Native DNA-Based Data Storage

DNA Data Storage Platforms

Native DNA-Based Computing
Our Readout Solution Summary

- **Select best reads for first alignment:** Best reads = highest quality addresses!
- **Perform alignment:** Roughly 30 traces (reads) involved.
- **Adjust sequence balance.** Balancing also limits runlengths!
- **Repeat while recruiting new traces.**
Deletion Correction Through Balancing

\[ C_{\text{est}} = \text{Current estimate of the consensus sequence} \]

- **Initial alignment:**
  - \( C_{\text{est}} \):
    \[
    \text{TTCACCCAAAAAACCAGAAACCAGCTTCAGCGA} \]
  - Trace1:
    \[
    \text{TTCACCCCAAACCGGAAACCCTTCAGCGA} \]
  - Trace2:
    \[
    \text{TTCACCCAAAAACCAGAAACCAGCTTCAGCGA} \]
  - Trace3:
    \[
    \text{TTCACCCAAAAACCAGAAACCAGCTTCAGCGA} \]

- **After MSA**
  - \( C'_{\text{est}} \):
    \[
    \text{T}^2\text{C}^1\text{A}^1\text{C}^4\text{A}^4 \ldots \]
  - Trace1:
    \[
    \text{T}^2\text{C}^1\text{A}^1\text{C}^3\text{A}^4 \ldots \]
  - Trace2:
    \[
    \text{T}^2\text{C}^1\text{A}^1\text{C}^4\text{A}^4 \ldots \]
  - Trace3:
    \[
    \text{T}^2\text{C}^1\text{A}^1\text{C}^3\text{A}^5 \ldots \]

- **After Balancing**
  - \( C'_{\text{est}} \):
    \[
    \text{T}^2\text{C}^1\text{A}^1\text{C}^3\text{A}^4 \ldots \]
  - Trace1:
    \[
    \text{T}^2\text{C}^1\text{A}^1\text{C}^3\text{A}^4 \ldots \]
  - Trace2:
    \[
    \text{T}^2\text{C}^1\text{A}^1\text{C}^3\text{A}^4 \ldots \]
  - Trace3:
    \[
    \text{T}^2\text{C}^1\text{A}^1\text{C}^3\text{A}^5 \ldots \]
Reading with Nanopores

**Consensus may still have errors:** Runlengths of As increase or decrease (protein-A interaction). Runlengths of Gs may form G quadruplexes.

**Example Statistics:**

<table>
<thead>
<tr>
<th>Block (length)</th>
<th>Number of reads</th>
<th>Sequencing Coverage depth</th>
<th>Number of errors: (substitution, insertion, deletion)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Per read (average)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>Maximum</td>
</tr>
<tr>
<td>1 (1,000)</td>
<td>201</td>
<td>176.145</td>
<td>192</td>
</tr>
<tr>
<td>2 (1,000)</td>
<td>407</td>
<td>315.521</td>
<td>349</td>
</tr>
<tr>
<td>3 (1,000)</td>
<td>490</td>
<td>460.375</td>
<td>482</td>
</tr>
<tr>
<td>4 (1,000)</td>
<td>100</td>
<td>81.763</td>
<td>87</td>
</tr>
<tr>
<td>5 (1,000)</td>
<td>728</td>
<td>688.663</td>
<td>716</td>
</tr>
<tr>
<td>6 (1,000)</td>
<td>136</td>
<td>120.907</td>
<td>129</td>
</tr>
<tr>
<td>7 (1,000)</td>
<td>577</td>
<td>542.78</td>
<td>566</td>
</tr>
<tr>
<td>8 (1,000)</td>
<td>217</td>
<td>199.018</td>
<td>207</td>
</tr>
<tr>
<td>9 (1,000)</td>
<td>86</td>
<td>56.828</td>
<td>75</td>
</tr>
<tr>
<td>10 (1,000)</td>
<td>442</td>
<td>396.742</td>
<td>427</td>
</tr>
<tr>
<td>11 (1,000)</td>
<td>114</td>
<td>101.826</td>
<td>110</td>
</tr>
<tr>
<td>12 (1,000)</td>
<td>174</td>
<td>162.559</td>
<td>169</td>
</tr>
<tr>
<td>13 (1,060)</td>
<td>378</td>
<td>352.35</td>
<td>366</td>
</tr>
<tr>
<td>14 (1,000)</td>
<td>222</td>
<td>189.918</td>
<td>203</td>
</tr>
<tr>
<td>15 (1,000)</td>
<td>236</td>
<td>222.967</td>
<td>232</td>
</tr>
<tr>
<td>16 (1,000)</td>
<td>198</td>
<td>182.99</td>
<td>195</td>
</tr>
<tr>
<td>17 (880)</td>
<td>254</td>
<td>240.273</td>
<td>250</td>
</tr>
</tbody>
</table>
Homopolymer Codes

> **Definition.** The integer sequence of a vector $x \in \mathbb{F}_4^n$ is the sequence of the length of the runs in $x$.

Example: $x = (0, 0, 1, 3, 3, 2, 1, 1) \rightarrow I(x) = (2, 1, 2, 1, 2)$.

The alternating sequence is the sequence of symbols in $x$, with all runs set to one. For above $x$, we have $S(x) = (0, 1, 3, 2, 1)$. 
Homopolymer Codes

- **Definition.** The integer sequence of a vector $x \in \mathbb{F}_4^n$ is the sequence of the length of the runs in $x$.

Example: $x = (0, 0, 1, 3, 3, 2, 1, 1) \rightarrow I(x) = (2, 1, 2, 1, 2)$.

The alternating sequence is the sequence of symbols in $x$, with all runs set to one. For above $x$, we have $S(x) = (0, 1, 3, 2, 1)$.

- Suppose that $C_H(n, t)$ is a code that can correct up to $t$ substitution errors. Let

$$C(n, t) = \{x \in \mathbb{F}_4^n : I(x) \mod 2 \in C_H(n, t)\}.$$ 

The code $C(n, t)$ can correct up to $t$ asymmetric (decreasing) run-preserving deletions.

- Related to sticky deletions [B90’s], [DA05].
Readout Time

- The sequencing time was $\sim 10$ hours, but only “junk” reads generated after $\sim 6$ hours.

- **How long shall we sequence for?**

- **Definition.** The $k$-deck of a sequence $x$ is the multiset of all subsequences of length $k$ of $x$.

- **Hybrid reconstruction:** One is given a small number $M$ of “long” asymmetric traces ($o(n)$ deletions). What is the smallest value of $k$ for a $k$-deck that along with the $M$ long traces ensures unique reconstruction?
Example images of Citizen Kane poster (1946) and Smiley. Only three deletions left after iterative alignment. Error-free decoding is possible with coding efficiency 88%
Native DNA-Based Data Storage
Resolving the Synthesis Problem?

- **Our system**: Store information in native DNA (e.g., *E. coli*). How?
- **Content cannot be changed easily**: –ATGCC– has to remain –ATGCC–.
Resolving the Synthesis Problem?

- **Our system**: Store information in native DNA (e.g., *E. coli*). How?
- **Content cannot be changed easily**: –ATGCC– has to remain –ATGCC–.
- **Structure? Symbol alphabet?**

![Cytosine and methylated Cytosine](image)

*Figure 1: Animation of nicked plasmid DNA. A nick can be enzymatically inserted or caused by shearing during plasmid preparation.*
Resolving the Synthesis Problem?

- **Our system:** Store information in native DNA (e.g., *E. coli*). How?
- **Content cannot be changed easily:** –ATGCC– has to remain –ATGCC–.
- **Structure? Symbol alphabet?**
- **Nicking using programmable DNA-guided artificial restriction enzymes** (in collaboration with Zhao lab).

![Diagram of Cytosine and methylated Cytosine](image)
Think of Punch Cards...

- **Our approach**: Do not nick or nick sense or antisense strand (ternary alphabet).
- **Reuse nicking enzyme on a large number of DNA “registers”**: PfAgo DNA-guided enzyme.
- **How do we know** that we have the “right nicks”?
How do we Read the Nicks?

- **Detect Nicks via Illumina Sequencers:** Denature nicked DNA, Sanger sequence (expensive).
How do we Read the Nicks?

- **Detect Nicks via Illumina Sequencers:** Denature nicked DNA, Sanger sequence (expensive).
- **Detect Nicks with Nanopores?** Collaboration with Radenovic lab, EPFL.
How do we Read the Nicks?

- **Detect Nicks via Illumina Sequencers:** Denature nicked DNA, Sanger sequence (expensive).
- **Simulate Nicks with Nanopores?** Collaboration with Leburton lab, EPFL.

Site of the Damage: T-T

- Molecule: 20 base-pair dsDNA (A-T)
- Concentration: 1 M KCl
- Ionic Voltage Bias: 1V
- Sheet Voltage Bias: 5mV
Sources of Errors

- **Erroneous Nicking**: Shifts in the positions of the nick (nicking window), missing nicks (deletions).
Sources of Errors

- **Erroneous Nicking:** Shifts in the positions of the nick (nicking window), missing nicks (deletions).
- **Erroneous Readout:** Shifts in the positions of the nick (nicking window), missing and inserted nicks (deletions and insertions).
Sources of Errors

- **Erroneous Nicking:** Shifts in the positions of the nick (nicking window), missing nicks (deletions).

- **Erroneous Readout:** Shifts in the positions of the nick (nicking window), missing and inserted nicks (deletions and insertions).

- **Instability Errors:** Do not want to nick one strand exclusively, as it may cause the backbone to break down.
New Coding Solutions I

- **Ternary Codes for Swap and Deletion correction**: Related to codes in the Damerau distance [GYM18].
- **Hard to Accommodate Instability Issues**: Revert to codewords described in terms of sets.
- **Formal Definition**: The *nicking codebook* is a set of subsets $S_i \subset [n]$, $i = 1, \ldots, M$, of fixed size $k$ such that for any $i \neq j$, one has $|S_i \cap S_j| \leq s$, where $M$ and $s$ are code design parameters.
New Coding Solutions I

- **Ternary Codes for Swap and Deletion correction**: Related to codes in the Damerau distance [GYM18].

- **Hard to Accommodate Instability Issues**: Revert to codewords described in terms of sets.

- **Formal Definition**: The *nicking codebook* is a set of subsets $S_i \subset [n]$, $i = 1, \ldots, M$, of fixed size $k$ such that for any $i \neq j$, one has $|S_i \cap S_j| \leq s$, where $M$ and $s$ are code design parameters.

- **Introduced by Babai and Frankl**: Let $q$ be a prime power and $1 \leq s \leq k \leq q$. Set $n = kq$. Let $\xi$ be a primitive element of $\mathbb{F}_q$ and $\mathcal{A} = \{0, 1, \xi, \ldots, \xi^{k-2}\}$. Then $|\mathcal{A}| = k$. For each polynomial $f \in \mathbb{F}_q[x]$, define

$$A_f := \{(a, f(a)) : a \in \mathcal{A}\}.$$

We also have $|A_f| = k$. Let

$$C(k, q, s) := \{A_f : f \in \mathbb{F}_q[x], \deg(f) \leq s - 1\}.$$

Then $C(k, q, s)$ is a collection of $q^s$ $k$-subsets of the set $X := \mathcal{A} \times \mathbb{F}_q$ and satisfies the property that every two sets intersect at at most $s - 1$ elements.
Set Discrepancy Theory

- Set discrepancy problem [Spencer’85, Lovasz’86]: Given a set of $m$ subsets \{A_1, \ldots, A_m\} of fixed size $k$ over a ground set $[n-1]$, find a labeling $\ell : [n-1] \rightarrow \{-1, +1\}$ which minimizes

\[
\max_{1 \leq i \leq m} \left| \sum_{x \in A_i} \ell(x) \right|, \ \text{i.e.}
\]

\[
\min_{\ell : [n-1] \rightarrow \{-1, +1\}} \max_{1 \leq i \leq m} \left| \sum_{x \in A_i} \ell(x) \right|.
\]

- In our case, the optimal mapping $\ell : [n-1] \rightarrow \{-1, +1\}$ ensures balance of nicks.
- Babai-Frankl sets can be shown to have zero discrepancy!
- Can extend the results further using Steiner systems.
Other Directions: Concentration Based Coding

- **Concentration based encoding**: Image processing in DNA.

![Diagram](image-url)
Other Directions: Concentration Based Coding

- **Concentration based encoding:** Image processing in DNA.
Other Directions: Concentration Based Coding

- **Concentration based encoding:** Image processing in DNA.
Other Directions: Enlarging the Code Alphabet

- **Enlarging the code alphabet**: Nonstructural, chemical modifications (with Schroeder lab).

- **Integration with nanoelectronics**: Changing random access approaches (with Li lab).
DNA Storage in Living Cells, Nature, 2017

- **Low-density storage using CRISPR-Cas, *E. coli***: Church et al., 2017.

  ![encoded GIF](image1)

  ![recalled GIF](image2)

- **Fountain DNA Storage**: Erlich et al., 2017 (Reed-Solomon in Grass et.al.: oligos treated as symbols over a large alphabet, redundancy at the oligo level).