Coding Techniques for Emerging DNA-Based Storage Systems

Olgica Milenkovic

A joint work with R. Gabrys, E. Garcia Ruiz, H. M. Kiah, J. Ma, G. J. Puleo, H. Tabatabaei, Y. Yuan, E. Yaakobi and H. Zhao

LIDS Student Conference, MIT

January 2016
The Future of Storage

- Cost of high-performance parallel storage: $0.3 per GB per month.

Figure: From Church, Harvard U.
The Future of Storage

- Cost of high-performance **parallel storage**: $0.3 per GB per month.
- Cost of **cloud storage**: Google Cloud $5.47 per 50 GB per month.

*Figure: From Church, Harvard U.*
The Future of Storage

- **Cost of high-performance parallel storage**: $0.3 per GB per month.
- **Cost of cloud storage**: Google Cloud $5.47 per 50 GB per month.
- **Cost of storage often minor compared to cost of access, processing, and data movement.**

*Figure: From Church, Harvard U.*
The Era of Massive Data

- **Large Hadron Collider**: 600 million collisions/s, 0.5 PB per week.

*Figure*: In search of the God particle, Wikipedia.
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.

Figure: In search of the God particle, Wikipedia.
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.

*Figure*: In search of the God particle, Wikipedia.
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 science (Twitter, Facebook, LinkedIn), NASA weather surveys, consumer and stock market data, Internet sources...

**Figure**: In search of the God particle, Wikipedia.
DNA as Storage Media

- **DNA is extremely durable:** Can still “read” mammoth, Neanderthal, 700,000 old horse DNA!
DNA as Storage Media

- **DNA is extremely durable**: Can still “read” mammoth, Neanderthal, 700,000 old horse DNA!
- **DNA write (synthesis) and read (sequencing) costs decrease daily.**
DNA as Storage Media

- **DNA is extremely durable:** Can still “read” mammoth, Neanderthal, 700,000 old horse DNA!
- **DNA write (synthesis) and read (sequencing) costs decrease daily.**
- **DNA information content of Human cell:** 6.4 GB. **Mass of a cell:** ~3 picograms. **No. of cells:** \(15 - 40 \times 10^{12}\).
DNA as Storage Media

- DNA is extremely durable: Can still “read” mammoth, Neanderthal, 700,000 old horse DNA!
- DNA write (synthesis) and read (sequencing) costs decrease daily.
- DNA information content of Human cell: 6.4 GB. Mass of a cell: ~ 3 picograms. No. of cells: 15 – 40 × 10^{12}.
- How much information can one store in a gram of DNA?
Implementations
Richard Feynman first to propose the use of macromolecules for storage ("There is plenty of room at the bottom").
Richard Feynman first to propose the use of macromolecules for storage ("There is plenty of room at the bottom").

Church et al. (Science, 2012) and Goldman et al. (Nature, 2013) stored 739 KB of data in synthetic DNA, mailed it and recreated the original digital files.
Richard Feynman first to propose the use of macromolecules for storage ("There is plenty of room at the bottom").

Church et al. (Science, 2012) and Goldman et al. (Nature, 2013) stored 739 KB of data in synthetic DNA, mailed it and recreated the original digital files.

Goal: a digital archival storage system that will safely store the equivalent of one million CDs in a gram of DNA for 10,000 years.
Renewed interest in DNA storage (UIUC, MS Research, IARPA Special Program on DNA-Based Storage).

**Goal:** Build a fully operational, cost-efficient, real-time, random access DNA-based memory.
Renewed interest in DNA storage (UIUC, MS Research, IARPA Special Program on DNA-Based Storage).

**Goal:** Build a fully operational, cost-efficient, real-time, random access DNA-based memory.

- Yazdi et.al., 2015 - First random access, rewritable DNA-based storage system. Encoded Wikipedia entries for six US universities (including MIT).
Our Experiments

- Random access achieved via **specialized address design**.
Our Experiments

- Random access achieved via **specialized address design**.
- Context identification and rewriting performed via **gBlock or OE-PCR methods**.

![Image of DNA gel and DNA synthesis processes](image)
The Write and Read Channels
The Write Channel: DNA Synthesis

Biochemistry of synthesis: Adding bases through deprotection & coupling cycles.
The Write Channel: DNA Synthesis

- DNA microarray based synthesis (left): Cost effective, large scale. Short strands, higher error rates.
The Write Channel: DNA Synthesis

- **DNA microarray based synthesis (left):** Cost effective, large scale. Short strands, higher error rates.
- **Long strand synthesis (right):** Synthesize via shorter blocks, assembled. Chemical error-correction.

![DNA microarray based synthesis](image)

![Long strand synthesis](image)
The Write Channel: DNA Synthesis

- **DNA microarray based synthesis (left):** Cost effective, large scale. Short strands, higher error rates.

- **Long strand synthesis (right):** Synthesize via shorter blocks, assembled. Chemical error-correction.

- Types of synthesis errors: predominantly substitutions, much less frequent deletions/insertions.
The Read Channel: Illumina and MinIon

- **Illumina (MiSeq, left):** Best overall performance of modern sequencing technologies in terms of yield and accuracy; large volumes of DNA reads, relatively small error rates (substitutions and context dependent deletions). Drawback short read length.
The Read Channel: Illumina and MinIon

- **Illumina (MiSeq, left):** Best overall performance of modern sequencing technologies in terms of yield and accuracy; large volumes of DNA reads, relatively small error rates (substitutions and context dependent deletions). Drawback short read length.

- **Oxford Nanopore - MinIon (Right):** Longer read length, miniaturized architecture. Large coverage errors, excessive number of block deletions.
The Read Channel: Shotgun Sequencing

Cloning /// Shearing /// Reading of unordered pool /// Computer aided alignment of overlapping fragments /// Consensus
Media Aging

- **Breakage - Bursty Deletions - Transpositions/Reversals:** No built-in mechanism for correcting damages.

<table>
<thead>
<tr>
<th>Type of Damage</th>
<th>Common Causes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double-strand break</td>
<td>Normal cellular activity, Ionizing radiation (including X-rays), Chemotherapeutic drugs, DNA repair of other types of damage</td>
</tr>
<tr>
<td>Chemical bond between neighboring nucleotides</td>
<td>Ultraviolet (UV) light</td>
</tr>
<tr>
<td>Chemical modification of a nucleotide</td>
<td>Reactive oxygen species (ROS), Chemotherapeutic drugs, Other cellular and environmental chemicals, Normal modifications that regulate what genes are active</td>
</tr>
<tr>
<td>Chemical Linkage of Two Strands</td>
<td>Reactive oxygen species (ROS), Chemotherapeutic drugs, Other cellular and environmental chemicals</td>
</tr>
</tbody>
</table>
Media Aging

- **Breakage - Bursty Deletions - Transpositions/Reversals:** No built-in mechanism for correcting damages.
- **Coupled with synthesis and sequencing errors...**

<table>
<thead>
<tr>
<th>Type of Damage</th>
<th>Common Causes</th>
</tr>
</thead>
</table>
| Double-strand break             | - Normal cellular activity  
                              | - Ionizing radiation (including X-rays)  
                              | - Chemotherapeutic drugs  
                              | - DNA repair of other types of damage |
| Chemical bond between neighboring nucleotides | - Ultraviolet (UV) light |
| Chemical modification of a nucleotide | - Reactive oxygen species (ROS)  
                              | - Chemotherapeutic drugs  
                              | - Other cellular and environmental chemicals  
                              | - Normal modifications that regulate what genes are active |
| Chemical Linkage of Two Strands | - Reactive oxygen species (ROS)  
                              | - Chemotherapeutic drugs  
                              | - Other cellular and environmental chemicals |
A formal mathematical theory of error-correction for DNA storage?
A formal mathematical theory of error-correction for DNA storage?
  - Microarray Synthesis and Shotgun Sequencing: DNA Profile Codes.
A formal mathematical theory of error-correction for DNA storage?

- Microarray Synthesis and Shotgun Sequencing: DNA Profile Codes.
- Microarray Synthesis and Nanopore Sequencing: Asymmetric Lee Distance Codes.
A formal mathematical theory of error-correction for DNA storage?

- Microarray Synthesis and Shotgun Sequencing: DNA Profile Codes.
- Microarray Synthesis and Nanopore Sequencing: Asymmetric Lee Distance Codes.
- DNA Media Aging: Codes in the Damerau Distance.
A formal **mathematical theory of error-correction** for DNA storage?
- Microarray Synthesis and Shotgun Sequencing: **DNA Profile Codes.**
- Microarray Synthesis and Nanopore Sequencing: **Asymmetric Lee Distance Codes.**
- DNA Media Aging: **Codes in the Damerau Distance.**
- Mathematical approaches for enabling random access and rewriting?
A formal **mathematical theory of error-correction** for DNA storage?
- Microarray Synthesis and Shotgun Sequencing: **DNA Profile Codes**.
- Microarray Synthesis and Nanopore Sequencing: **Asymmetric Lee Distance Codes**.
- DNA Media Aging: **Codes in the Damerau Distance**.

Mathematical approaches for enabling random access and rewriting?
- Address Design: **(Weakly) Mutually Uncorrelated Codes**.
A formal mathematical theory of error-correction for DNA storage?
  - Microarray Synthesis and Shotgun Sequencing: DNA Profile Codes.
  - Microarray Synthesis and Nanopore Sequencing: Asymmetric Lee Distance Codes.
  - DNA Media Aging: Codes in the Damerau Distance.

Mathematical approaches for enabling random access and rewriting?
  - Address Design: (Weakly) Mutually Uncorrelated Codes.
  - Controlled Assembly: Uncorrelated Array Codes.
A formal mathematical theory of error-correction for DNA storage?
A formal mathematical theory of error-correction for DNA storage?

- Microarray Synthesis and Nanopore Sequencing: **Asymmetric Lee Distance (ALD) Codes.**
A formal mathematical theory of error-correction for DNA storage?

Microarray Synthesis and Nanopore Sequencing: Asymmetric Lee Distance (ALD) Codes.

For a positive integer $\lambda$, the ALD $d_\lambda((a; b), (c; d))$ between pairs of binary sequences $(a; b), (c; d)$ is defined as:

$$d_\lambda((a; b), (c; d)) = \sum_{i=1}^{n} (1 + \lambda) (\mathbb{1}(a_i, b_i) + \mathbb{1}(c_i, d_i)) + \lambda \mathbb{1}(a_i, \overline{b_i}, \overline{c_i}, d_i) - 2(1 + \lambda) \mathbb{1}(a_i, b_i, c_i, d_i).$$
• A formal **mathematical theory of error-correction** for DNA storage?
  • DNA Media Aging: **Codes in the Damerau Distance.**
A formal **mathematical theory of error-correction** for DNA storage?

- DNA Media Aging: **Codes in the Damerau Distance.**
- The **Damerau–Levenshtein (DL) distance** is a string metric, which for two strings over a finite alphabet equals the minimum number of insertions, deletions, substitutions and adjacent transpositions needed to transform one string into the other.
A formal mathematical theory of error-correction for DNA storage?

- DNA Media Aging: Codes in the Damerau Distance.
- The Damerau–Levenshtein (DL) distance is a string metric, which for two strings over a finite alphabet equals the minimum number of insertions, deletions, substitutions and adjacent transpositions needed to transform one string into the other.
- The block DL distance: Extension in which edit units are blocks of limited length.
• A formal mathematical theory of error-correction for DNA storage?
  • DNA Media Aging: Codes in the Damerau Distance.
  • The Damerau–Levenshtein (DL) distance is a string metric, which for two strings over a finite alphabet equals the minimum number of insertions, deletions, substitutions and adjacent transpositions needed to transform one string into the other.
  • The block DL distance: Extension in which edit units are blocks of limited length.
  • Varshamov-Tenengoltz’s extensions for the DL distance: Uses the derivative of $a, a' = (a_1, a_2 + a_1, a_3 + a_2, \ldots, a_n + a_{n-1})$. 


A formal mathematical theory of error-correction for DNA storage?

- DNA Media Aging: Codes in the Damerau Distance.
- The Damerau–Levenshtein (DL) distance is a string metric, which for two strings over a finite alphabet equals the minimum number of insertions, deletions, substitutions and adjacent transpositions needed to transform one string into the other.
- The block DL distance: Extension in which edit units are blocks of limited length.
- Varshamov-Tenengoltz's extensions for the DL distance: Uses the derivative of \( a, a' = (a_1, a_2 + a_1, a_3 + a_2, \ldots, a_n + a_{n-1}) \).
- Component codes: \( C_H(n, 3) \) a single error-correcting code; \( C_D(n) \) a single deletion-correcting code.

\[
C_{TVD}(n) = \{ a \in \mathbb{F}_2^n : a \in C_D(n), a' \in C_H(n, 3) \}.
\]

The code \( C_{TVD}(n) \) can correct one single deletion or adjacent transposition.
Mathematical approaches for enabling random access and rewriting?
- Address Design: (Weakly) Mutually Uncorrelated Codes.
A formal mathematical theory of error-correction for DNA storage?
  • **Microarray Synthesis and Shotgun Sequencing:** DNA Profile Codes.
  • **Microarray Synthesis and Nanopore Sequencing:** Asymmetric Lee Distance Codes.
  • **DNA Media Aging:** Codes in the Damerau Distance.

Mathematical approaches for enabling random access and rewriting?
  • **Address Design:** (Weakly) Mutually Uncorrelated Codes.
  • **Controlled Assembly:** Uncorrelated Array Codes.
The DNA Storage Channel
DNA Storage Channel: Basics

Synthesis channel captures the “write” process.
The sequence synthesis process introduces errors (current technologies $\leq 0.1\%$).
DNA sequencing represents the “read” process. Consists of fragmenting sequence to be read, and “reconstructing” fragments (\(\ell\)-grams).
DNA Storage Channel: Basics

Note that strings at the output of the fragmentation block are *not* ordered.
DNA Storage Channel: Basics

- Sequencing introduces errors in some $\ell$-grams and some $\ell$-grams may not be covered.
DNA Storage Channel: Basics

- Sequencing introduces errors in some \( \ell \)-grams and some \( \ell \)-grams may not be covered.
- Modern Illumina platforms have substitution error rates \( \leq 0.5\% \). Coverage errors context-dependent.
DNA Storage Channel: Profile Vectors

Profile vectors

The profile vector of a sequence reflects the count of its $\ell$-grams;
**DNA Storage Channel: Profile Vectors**

Profile vectors
The profile vector of a sequence reflects the count of its \( \ell \)-grams;

**Example**
Profile of vector \( \mathbf{x} = 10011001 \) equals

\[
\begin{array}{cccccccc}
000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\
0 & 2 & 0 & 1 & 2 & 0 & 1 & 0
\end{array}
\]
Profile Vectors: Formal Definition

Fix alphabet size $q$ and fragment (read) length $\ell < n$. The profile vector of some sequence $x$, denoted by $p(x; q, \ell)$, has length $q^\ell$ and its entry indexed by $z$ equals the number of occurrences of $z$ in $x$ as an $\ell$-gram.
Input and Output Profile Vectors

Example
Profile of \(x = 10011001\) and sequencing channel output:

\[
\begin{array}{ccccccccccc}
000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\
(0, & 2, & 0, & 1, & 2, & 0, & 1, & 0), \\
(0, & 1, & 0, & 2, & 0, & 1, & 1, & 0). \\
\end{array}
\]
Motivation

Implementations

The Write and Read Channels

The DNA Storage Channel

Profile Vectors and $\ell$-Gram Codes

THANK YOU!

Code Design Criteria

<table>
<thead>
<tr>
<th>Message</th>
<th>Encoded Codeword</th>
<th>DNA Storage Channel</th>
<th>Output Profile Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes-No</td>
<td>{ 10011001 }</td>
<td>y = 10101010</td>
<td>p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0)</td>
</tr>
</tbody>
</table>

Condition 1: Codewords should have profile vectors that are sufficiently "distinct," i.e., one should be able to correct combination of synthesis substitution (burst), coverage, and $\ell$-gram errors.

Definition 1: The $\ell$-gram distance between $x$ and $y$ equals the asymmetric distance $p(x; q, \ell)$ between $p(x; q, \ell)$ and $p(y; q, \ell)$.

$p(x; 2, 3) = (0, 2, 0, 1, 2, 0, 1, 0)$

$p(y; 2, 3) = (0, 0, 3, 0, 0, 3, 0, 0)$

$p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0)$. 
## Code Design Criteria

<table>
<thead>
<tr>
<th>Message</th>
<th>Codeword</th>
<th>DNA Storage Channel</th>
<th>Output profile vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>(x = 10011001)</td>
<td></td>
<td>(p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0))</td>
</tr>
<tr>
<td>No</td>
<td>(y = 10101010)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- \(p(x; 2, 3) = (0, 2, 0, 1, 2, 0, 1, 0)\)
- \(p(y; 2, 3) = (0, 0, 3, 0, 0, 3, 0, 0)\)
- \(p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0)\).

### Condition 1

Codewords should have **profile vectors** that are sufficiently “distinct,” i.e., one should be able to correct combination of **synthesis substitution (burst)**, **coverage**, and \(\ell\)-**gram errors.**
Code Design Criteria

Message \(\rightarrow\) Encoder \(\rightarrow\) Codeword \(\rightarrow\) DNA Storage Channel \(\rightarrow\) Output profile vector

\{ Yes, No \} \{ \(x = 10011001\) \(y = 10101010\) \}

\(p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0)\)

\begin{align*}
p(x; 2, 3) & = (0, 2, 0, 1, 2, 0, 1, 0) \\
p(y; 2, 3) & = (0, 0, 3, 0, 3, 0, 0) \\
p(\hat{x}; 2, 3) & = (0, 1, 0, 2, 0, 1, 1, 0)
\end{align*}

**Condition 1**

Codewords should have **profile vectors** that are sufficiently “distinct,” i.e., one should be able to correct combination of synthesis substitution (burst), coverage, and \(\ell\)-gram errors.

**Definition 1**

The \(\ell\)-gram distance between \(x\) and \(y\) equals the asymmetric distance (\(\ell_1\) distance) between \(p(x; q, \ell)\) and \(p(y; q, \ell)\).
Code Design Criteria

Encoder

```
Message

{ Yes }

{ No }  

{ x = 10011001

y = 10101010

}  

```

Codeword

DNA Storage Channel

Output profile vector

```
p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0)
```

```
p(x; 2, 3) = (0, 2, 0, 1, 2, 0, 1, 0)
p(y; 2, 3) = (0, 0, 3, 0, 0, 3, 0, 0)
p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0).
```

Definition 1

The \( \ell \)-gram distance between \( x \) and \( y \) equals the asymmetric distance (\( \ell_1 \) distance) between \( p(x; q, \ell) \) and \( p(y; q, \ell) \).
Code Design Criteria

\[
\begin{array}{cccc}
\text{Message} & \text{Encoder} & \text{Codeword} & \text{DNA Storage Channel} \\
\{ \text{Yes} \} & \{ x = 10011001 \} & \{ y = 10101010 \} & p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0) \\
\{ \text{No} \} & p(x; 2, 3) = (0, 2, 0, 1, 2, 0, 1, 0) \\
 & p(y; 2, 3) = (0, 0, 3, 0, 3, 0, 0) \\
 & p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0).
\end{array}
\]

**Definition 1**

The $\ell$-gram distance between $x$ and $y$ equals the asymmetric distance ($\ell_1$ distance) between $p(x; q, \ell)$ and $p(y; q, \ell)$.

**Asymmetric Distance**

Let $u, v \in \mathbb{Z}_{\geq 0}^N$. Define $\Delta(u, v) = \sum_{i=1}^N \max(u_i - v_i, 0)$. 

Asymmetric distance: $d_{\text{asym}}(u, v) = \max(\Delta(u, v), \Delta(v, u))$. 
Asymmetric Distance

Let \( u, v \in \mathbb{Z}_{\geq 0}^N \). Define

\[
\Delta(u, v) = \sum_{i=1}^{N} \max(u_i - v_i, 0).
\]

**Asymmetric distance:**

\[
d_{\text{asym}}(u, v) = \max(\Delta(u, v), \Delta(v, u)).
\]
Motivation
Implementations
The Write and Read Channels
The DNA Storage Channel
Profile Vectors and $\ell$-Gram Codes
THANK YOU!

Code Design Criteria

<table>
<thead>
<tr>
<th>p(x; 2, 3)</th>
<th>000</th>
<th>001</th>
<th>010</th>
<th>011</th>
<th>100</th>
<th>101</th>
<th>110</th>
<th>111</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(y; 2, 3)</td>
<td>(0, 2, 0, 1, 2, 0, 1, 0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p(\hat{x}; 2, 3)</td>
<td>(0, 1, 0, 2, 0, 1, 1, 0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Asymmetric Distance

Let $u, v \in \mathbb{Z}_0^N$. Define $\Delta(u, v) = \sum_{i=1}^{N} \max(u_i - v_i, 0)$.

Asymmetric distance: $d_{\text{asym}}(u, v) = \max(\Delta(u, v), \Delta(v, u))$.

Minimum Asymmetric Distance

A DNA storage code with minimum asymmetric distance $d$ can correct $s_1$ substitution errors due to synthesis, $s_2$ substitution errors due to sequencing and $t$ coverage errors provided that $d > 2s_1 + 2s_2 + t$. 
**Code Design Criteria**

Message → **Encoder** → **Codeword** → **DNA Storage Channel** → Output profile vector

\[
\begin{align*}
\text{Encoder:} & \quad \{ \text{Yes} \} \quad \{ \text{No} \} \\
\text{Codeword:} & \quad \{ x = 10011001 \} \quad \{ y = 10101010 \}
\end{align*}
\]

\[p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0)\]

\[p(x; 2, 3) = (0, 2, 0, 1, 2, 0, 1, 0)\]

\[p(y; 2, 3) = (0, 0, 3, 0, 0, 3, 0, 0)\]
## Code Design Criteria

- **Motivation**
- **Implementations**
- **The Write and Read Channels**
- **The DNA Storage Channel**
- **Profile Vectors and $\ell$-Gram Codes**
- **THANK YOU!**

### Code Design Criteria

<table>
<thead>
<tr>
<th>Message</th>
<th>Codeword</th>
<th>DNA Storage Channel</th>
<th>Output profile vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>{ Yes, No }</td>
<td>$x = 10011001$, $y = 10101010$</td>
<td></td>
<td>$p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0)$</td>
</tr>
</tbody>
</table>

**Condition 2**

Codewords whose $\ell$-grams avoid error-causing substrings.

<table>
<thead>
<tr>
<th>$p(x; 2, 3)$</th>
<th>000</th>
<th>001</th>
<th>010</th>
<th>011</th>
<th>100</th>
<th>101</th>
<th>110</th>
<th>111</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$p(y; 2, 3)$</th>
<th>000</th>
<th>001</th>
<th>010</th>
<th>011</th>
<th>100</th>
<th>101</th>
<th>110</th>
<th>111</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Code Design Criteria

Message → Encoder → Codeword → DNA Storage Channel → Output profile vector

Encodings:

- \( x = 10011001 \)
- \( y = 10101010 \)

Output profile vector:

- \( p(\hat{x}; 2, 3) = (0, 1, 0, 2, 0, 1, 1, 0) \)
- \( p(\hat{y}; 2, 3) = (0, 0, 3, 0, 0, 3, 0, 0) \)

Condition 2

Codewords whose \( \ell \)-grams avoid error-causing substrings.

Avoid “bad” grams that cause sequencing errors and media instability:

- Weight profiles of \( \ell \)-grams. GC content roughly 50%.
- Forbidden \( \ell \)-grams. Certain substrings, such as \( GCG \) and \( CGC \), are likely to cause coverage errors.
**Code Design Criteria**

![Diagram](image)

**Condition 2**

Codewords whose \( \ell \)-grams avoid error-causing substrings.

Avoid “bad” grams that cause sequencing errors and media instability:

- **Weight profiles of \( \ell \)-grams.** GC content roughly 50%.
- **Forbidden \( \ell \)-grams.** Certain substrings, such as \( GCG \) and \( CGC \), are likely to cause coverage errors.

For example, may require that \( \ell \)-grams lie in \( S = \{001, 010, 011, 100, 101, 110\} \).
Distinct $\ell$-gram Profile Vectors

Let $Q(n; S)$ be the largest set of $q$-ary words of length $n$ whose $\ell$-grams belong to $S$, and which have distinct $\ell$-gram profile vectors.

Determine the size of $Q(n; S)$. 

Fundamental Questions

\( \ell \)-gram Reconstruction Code (GRC)

\( C \subseteq Q(n; S') \) is an \((n, d; S')\)-\( \ell \)-GRC if the \( \ell \)-gram distance between any pair of distinct words is at least \( d \).

**Construct good \((n, d; S')\)-\( \ell \)-GRC.** "Good" means large codebook size, avoidance of bad \( \ell \)-grams.
Profile Vectors and ℓ-Gram Codes
Example for $q = 2$, $\ell = 3$.

Nodes are $q$-ary strings of length $\ell - 1$. $(v, v')$ is an arc if

$$
\begin{align*}
&v_2 \quad v_3 \quad v_{\ell-1} \\
&\| \quad \| \quad \ldots \quad \| \\
&v'_1 \quad v'_2 \quad v'_{\ell-2}
\end{align*}
$$
Let $S(\ell; w_1, w_2)$ denote the binary strings of length $\ell$ with weight between $w_1$ and $w_2$. 

**Restricted De Bruijn Graphs**

Nodes $V$ are $\ell-1$-prefixes and -suffixes of strings in $S$. 

$(v, v')$ is an arc if

\[
\begin{align*}
v_2 & \parallel v_3 & \parallel \cdots & \parallel v_{\ell-1} \\
v'_1 & \parallel v'_2 & \parallel & \cdots & \parallel v'_{\ell-2} \\
\end{align*}
\]

and $v_1 v_2 \cdots v_{\ell-1} v'_{\ell-1} \in S$. 

**Restricted De Bruijn Graphs $D(S)$**

Let $S = S(3; 1, 2)$ and $S = S(4; 2, 3)$.
Profile Vectors and Flows

Representation of profile vectors of words in $Q(n; S)$ using the digraph $D(S)$. 
Profile Vectors and Flows

The representation of profile vectors of words in $Q(n; S)$ using the digraph $D(S)$.

Closed Words

Closed words are words that start and end with the same $(\ell - 1)$-gram. $\overline{Q}(n; S)$: largest set of $q$-ary closed words of length $n$ whose $\ell$-grams belong to $S$ and which have distinct $\ell$-gram profiles.
Profile Vectors and Flows

Representation of profile vectors of words in $Q(n; S)$ using the digraph $D(S)$.

Closed Words

Closed words are words that start and end with the same $(\ell - 1)$-gram.

$Q(n; S)$: largest set of $q$-ary closed words of length $n$ whose $\ell$-grams belong to $S$ and which have distinct $\ell$-gram profiles.

Flows

Paths in $D(S)$ such that sum of incoming arc weights is equal to sum of outgoing arc weights at each vertex. Profile vectors of words in $\overline{Q}(n; S)$ are flow vectors in $D(S)$. 
Necessary Conditions

Let \( \mathbf{u} \) be a profile vector (of a closed word). Then \( \mathbf{u} \) satisfies the following conditions.

Flow conservations equations:

\[
\mathbf{B}\mathbf{u} = \mathbf{0},
\]

where \( \mathbf{B} \) be the incidence matrix of \( D(S) \).

Sum of flows:

\[
1\mathbf{u} = n - \ell + 1.
\]

Let \( \mathbf{A} = \begin{pmatrix} \mathbf{1} \\ \mathbf{B} \end{pmatrix} \) and \( \mathbf{b} = (1, 0, \ldots, 0)^T \). Rewrite equations as

\[
\mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b} \quad \text{and} \quad \mathbf{u} \geq 0.
\]
Sufficient Conditions

Flows are not always profile vectors

Let \( u \geq 0 \) be such that

\[
Au = (n - \ell + 1)b.
\]

This does not imply that \( u \) is a profile vector!
Sufficient Conditions

If all flows are **positive**, then the flow vector is indeed a profile vector.

Profile vector of \textbf{0110101110011}.

\[ Au = (n - \ell + 1)b \text{ and } u > 0. \]
Profile Vectors and Lattice Points

Consider the following two sets of lattice points:

\[ F(n; S) = \{ u \in \mathbb{Z}^{|S|} : Au = (n - \ell + 1)b, \ u \geq 0 \}, \]
\[ E(n; S) = \{ u \in \mathbb{Z}^{|S|} : Au = (n - \ell + 1)b, \ u > 0 \}. \]

Clearly, one has

\[ |E(n; S)| \leq |Q(n; S)| \leq |F(n; S)|. \]
Profile Vectors and Lattice Points

Consider the following two sets of lattice points:

\[ \mathcal{F}(n; S) = \{ u \in \mathbb{Z}^{|S|} : Au = (n - \ell + 1)b, \ u \geq 0 \}, \]
\[ \mathcal{E}(n; S) = \{ u \in \mathbb{Z}^{|S|} : Au = (n - \ell + 1)b, \ u > 0 \}. \]

Clearly, one has

\[ |\mathcal{E}(n; S)| \leq |\mathcal{Q}(n; S)| \leq |\mathcal{F}(n; S)|. \]

Observations

- \( \mathcal{F}(n; S) \) is a polytope. It can be shown to be of dimension \(|S| - |V(S)|\).
- \( \mathcal{E}(n; S) \) is the interior of \( \mathcal{F}(n; S) \) if \( D(S) \) is strongly connected.
- May use Ehrhart theory for polytopes to determine \(|\mathcal{E}(n; S)|, |\mathcal{F}(n; S)|.\)
Consider the following two sets of lattice points:

\[ \mathcal{F}(n; S) \triangleq \{ u \in \mathbb{Z}^{\lvert S \rvert} : Au = (n - \ell + 1)b, \ u \geq 0 \}, \]
\[ \mathcal{E}(n; S) \triangleq \{ u \in \mathbb{Z}^{\lvert S \rvert} : Au = (n - \ell + 1)b, \ u > 0 \}. \]

\[ |\mathcal{E}(n; S)| \leq |\mathcal{Q}(n; S)| \leq |\mathcal{F}(n; S)|. \]
Profile Vectors and Lattice Points

Consider the following two sets of lattice points:

\[ \mathcal{F}(n; S) \triangleq \{ u \in \mathbb{Z}^{|S|} : Au = (n - \ell + 1)b, \ u \geq 0 \}, \]
\[ \mathcal{E}(n; S) \triangleq \{ u \in \mathbb{Z}^{|S|} : Au = (n - \ell + 1)b, \ u > 0 \}. \]

\[ |\mathcal{E}(n; S)| \leq |\mathcal{Q}(n; S)| \leq |\mathcal{F}(n; S)|. \]

Observations

- Define the polytope
  \[ P_S = \{ u \in \mathbb{R}^{|S|} : Au = b, \ u \geq 0 \}. \]
- \( \mathcal{F}(n; S) \) is the set of lattice points in \( (n - \ell + 1)P_S \).
- \( \mathcal{E}(n; S) \) is the set of lattice points in the interior of \( (n - \ell + 1)P_S \).
For a polytope $\mathcal{P} \subset \mathbb{R}^N$ and $t \in \mathbb{R}$, the dilation $t\mathcal{P}$ is given by

$$t\mathcal{P} = \{tx : x \in \mathcal{P}\}.$$ 

The lattice point enumerator for $\mathcal{P}$ is $L_\mathcal{P} : \mathbb{R} \to \mathbb{Z}$ defined by

$$L_\mathcal{P}(t) = |t\mathcal{P} \cap \mathbb{Z}^N|.$$
Lattice Point Enumeration in Dilated Polytopes

For a polytope $P \subset \mathbb{R}^N$ and $t \in \mathbb{R}$, the dilation $tP$ is given by
$$tP = \{tx : x \in P\}.$$ 

The lattice point enumerator for $P$ is $L_P : \mathbb{R} \to \mathbb{Z}$ defined by
$$L_P(t) = |tP \cap \mathbb{Z}^N|.$$

**Theorem (Ehrhart)**

If $P$ is a rational $D$-dimensional polytope, then $L_P(t)$ is a “quasipolynomial” (polynomial with periodic functions as coefficients) in $t$ of degree $D$. 
For a polytope $\mathcal{P} \subset \mathbb{R}^N$ and $t \in \mathbb{R}$, the dilation $t\mathcal{P}$ is given by
\[ t\mathcal{P} = \{tx : x \in \mathcal{P}\}. \]

The lattice point enumerator for $\mathcal{P}$ is $\mathcal{L}_{\mathcal{P}} : \mathbb{R} \rightarrow \mathbb{Z}$ defined by
\[ \mathcal{L}_{\mathcal{P}}(t) = |t\mathcal{P} \cap \mathbb{Z}^N|. \]
Lattice Point Enumeration in Dilated Polytopes

For a polytope $\mathcal{P} \subset \mathbb{R}^N$ and $t \in \mathbb{R}$, the dilation $t\mathcal{P}$ is given by

$$t\mathcal{P} = \{tx : x \in \mathcal{P}\}.$$ 

The lattice point enumerator for $\mathcal{P}$ is $\mathcal{L}_{\mathcal{P}} : \mathbb{R} \to \mathbb{Z}$ defined by

$$\mathcal{L}_{\mathcal{P}}(t) = |t\mathcal{P} \cap \mathbb{Z}^N|.$$ 

**Theorem (Ehrhart-Macdonald’s Reciprocity)**

The number of lattice points in the interior of $t\mathcal{P}$ is given by $(−1)^D \mathcal{L}_{\mathcal{P}}(−t)$, and is thus a “quasipolynomial” of degree $D$. 

---

Motivation  Implementations  The Write and Read Channels  The DNA Storage Channel  Profile Vectors and ℓ-Gram Codes  THANK YOU!
Main Enumeration Results

**Theorem**

Suppose \( D(S) \) is strongly connected. Then \( |\mathcal{E}(n; S)| \) and \( |\mathcal{F}(n; S)| \) are both quasipolynomials in \( n \) of the same degree \( |S| - |V(S)| \). In particular,

\[
|\overline{Q}(n; S)| = \Theta'(n^{|S| - |V(S)|})
\]
Main Enumeration Results

Theorem

Suppose $D(S)$ is strongly connected. Then $|\mathcal{E}(n; S)|$ and $|\mathcal{F}(n; S)|$ are both quasipolynomials in $n$ of the same degree $|S| - |V(S)|$. In particular, $|\overline{Q}(n; S)| = \Theta'(n^{|S| - |V(S)|})$.

Here, $|\overline{Q}(n; S)| = \Theta'(n^3)$. 
Corollaries of Main Enumeration Result

Here, $|\overline{Q}(n; S)| = \frac{n^3}{288} + O(n^2)$ (Curtesy of Latte).

Theorem (Jacquet, Knessl, Szpankowski, 2012; Ukkonnen, Pevzner 1990’s)

Fix $q, \ell$ and let $S$ be the set of all $q$-ary strings of length $\ell$. Then

\[ |\mathcal{E}(n; S)| \sim |\mathcal{F}(n; S)| \sim |\overline{Q}(n; S)| \sim c(S)n^{q\ell - q^{\ell-1}} \]

where $c(S)$ is a constant.

$f \sim g$ means that $\lim_{n \to \infty} f(n)/g(n) = 1$.

Corollary

Suppose $D(S)$ is strongly connected and contains loops. Then

\[ |\mathcal{E}(n; S)| \sim |\mathcal{F}(n; S)| \sim |\overline{Q}(n; S)| \sim c(S)n^{|S| - |V|} \]

where $c(S)$ is a constant.
Fix $d$ and let $p$ be a prime such that $p > d$ and $p > N$. Choose $N$ distinct nonzero elements $\alpha_1, \alpha_2, \ldots, \alpha_N$ in $\mathbb{Z}/p\mathbb{Z}$ and consider the matrix

$$H = \begin{pmatrix}
\alpha_1 & \alpha_2 & \cdots & \alpha_N \\
\alpha_1^2 & \alpha_2^2 & \cdots & \alpha_N^2 \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_1^d & \alpha_2^d & \cdots & \alpha_N^d
\end{pmatrix}.$$

Pick any vector $\beta \in (\mathbb{Z}/p\mathbb{Z})^N$ and define the code

$$C(H, \beta) = \{ u \in \mathbb{Z}^N : Hu \equiv \beta \mod p \}.$$
Fix $d$ and let $p$ be a prime such that $p > d$ and $p > N$. Choose $N$ distinct nonzero elements $\alpha_1, \alpha_2, \ldots, \alpha_N$ in $\mathbb{Z}/p\mathbb{Z}$ and consider the matrix

$$
H = \begin{pmatrix}
\alpha_1 & \alpha_2 & \cdots & \alpha_N \\
\alpha_1^2 & \alpha_2^2 & \cdots & \alpha_N^2 \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_1^d & \alpha_2^d & \cdots & \alpha_N^d
\end{pmatrix}.
$$

Pick any vector $\beta \in (\mathbb{Z}/p\mathbb{Z})^N$ and define the code

$$
C(H, \beta) = \{ u \in \mathbb{Z}^N : Hu \equiv \beta \mod p \}.
$$

**Theorem (Varshamov, 1973)**

$C(H, \beta)$ is a code with minimum asymmetric distance $d + 1$. 

THANK YOU!
Gram Reconstruction Codes

Construction 1

Let \( pQ(n; S) \) be the set of distinct profile vectors of words in \( S \) and \( N = |S| \). Then \( C(H, \beta) \cap pQ(n; S) \) is an \((n, d + 1; S)\)-\(\ell\)-gram reconstruction code.
Gram Reconstruction Codes

Construction I

Let $pQ(n; S)$ be the set of distinct profile vectors of words in $S$ and $N = |S|$. Then $C(H, \beta) \cap pQ(n; S)$ is an $(n, d + 1; S)$-$\ell$-gram reconstruction code.

Example

Let $q = 2$, $\ell = 3$, $S = \{001, 010, 011, 100, 101, 110\}$ and so, $N = 6$. Let $d = 3$ and pick $p = 7$, so that

$$H = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 4 & 2 & 2 & 4 & 1 \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Then $C(H, \beta)$ contains the following words.

<table>
<thead>
<tr>
<th>$(4, 0, 0, 1, 0, 1)$</th>
<th>$(0, 1, 1, 4, 0, 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(2, 2, 0, 2, 0, 0)$</td>
<td>$\leftrightarrow$</td>
</tr>
<tr>
<td>$(1, 4, 0, 0, 1, 0)$</td>
<td>$\leftrightarrow$</td>
</tr>
<tr>
<td>$(1, 1, 1, 1, 1, 1)$</td>
<td>$\leftrightarrow$</td>
</tr>
<tr>
<td>$(1, 0, 1, 0, 0, 4)$</td>
<td></td>
</tr>
</tbody>
</table>

Above, three profile vectors in $pQ(8; S)$. 
Gram Reconstruction Codes

**Construction 1**

Let $pQ(n; S)$ be the set of distinct profile vectors of words in $S$ and $N = |S|$. Then $C(H, \beta) \cap pQ(n; S)$ is an $(n, d + 1; S)$-$\ell$-gram reconstruction code.
Gram Reconstruction Codes

Construction I

Let $pQ(n; S)$ be the set of distinct profile vectors of words in $S$ and $N = |S|$. Then $C(H, \beta) \cap pQ(n; S)$ is an $(n, d + 1; S)$-$\ell$-gram reconstruction code.

Pigeonhole principle: there exists a $\beta$ such that $|C(H, \beta) \cap pQ(n; S)|$ is at least $|pQ(n; S)|/p^d$.

However, the optimal choice of $\beta$ is not known.
Construction 1

Let $pQ(n; S)$ be the set of distinct profile vectors of words in $S$ and $N = |S|$. Then $C(H, \beta) \cap pQ(n; S)$ is an $(n, d + 1; S)$-\(\ell\)-gram reconstruction code.

Pigeonhole principle: there exists a $\beta$ such that $|C(H, \beta) \cap pQ(n; S)|$ is at least $|pQ(n; S)|/p^d$.

However, the optimal choice of $\beta$ is not known.

We fix a certain choice of $H$ and $\beta$ and provide lower bounds on the size of $C(H, \beta) \cap pQ(n; S)$ as a function of $n$. 
The Write and Read Channels

The DNA Storage Channel

Profile Vectors and $\ell$-Gram Codes

THANK YOU!

Ehrhart Theory Continued

Define the $(|V| + 1 + d) \times (|S| + d)$-matrix

$$A_{\text{GRC}} = \begin{pmatrix} A & 0 \\ H & -pI_d \end{pmatrix}.$$ 

Proposition

If $D(S)$ is strongly connected and $\mathcal{C}(H, 0) \cap \text{Null}_{>0}B$ is nonempty, then $|\mathcal{C}(H, 0) \cap pQ(n; S)|$ is at least the number of lattice points in the interior of the polytope

$$\left\{ u \in \mathbb{R}^{|S|+d} : A_{\text{GRC}}u = (n - \ell + 1)b \right\}.$$

- Null$_{>0}B$ denotes the set of vectors in the null space of $B$ with strictly positive entries.
Ehrhart Theory Continued

Define the \((|V| + 1 + d) \times (|S| + d)\)-matrix

\[ A_{\text{GRC}} = \begin{pmatrix} A & 0 \\ H & -pI_d \end{pmatrix}. \]

**Proposition**

If \(D(S)\) is strongly connected and \(C(H, 0) \cap \text{Null}_{>0} B\) is nonempty, then

\[ |C(H, 0) \cap \mathfrak{p}Q(n; S)| \text{ is at least the number of lattice points in the interior of} \]

\[ \left\{ u \in \mathbb{R}^{|S|+d} : A_{\text{GRC}} u = (n - \ell + 1)b \right\}. \]

- \(\text{Null}_{>0} B\) denotes the set of vectors in the null space of \(B\) with strictly positive entries.

**Theorem**

If \(D(S)\) is strongly connected and \(C(H, 0) \cap \text{Null}_{>0} B\) is nonempty, then

\[ |C(H, 0) \cap \mathfrak{p}Q(n; S)| = \Omega' \left( n^{|S| - |V(S)|} \right). \]

- \(f(n) = \Omega'(g(n))\) means that for a fixed value of \(\ell\), there exists an integer \(\lambda\) and a positive constant \(c\) so that \(f(n) \geq cg(n)\) for sufficiently large \(n\) with \(\lambda|n - \ell + 1|).
Encoding and Decoding?

- **Encoding**: Systematic encoder that takes profile of input $x$ and converts it into redundant profile.
Encoding and Decoding?

- **Encoding**: Systematic encoder that takes profile of input $x$ and converts it into redundant profile.

- **Decoding**: Receive profile. Correct errors in profile. Assemble profile (say, by using Hierholzer’s algorithm).


THANK YOU!