

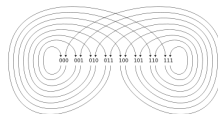
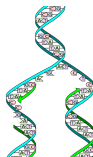
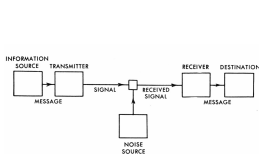
Coding Techniques for Emerging DNA-Based Storage Systems

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



Motivation

The Future of Storage

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

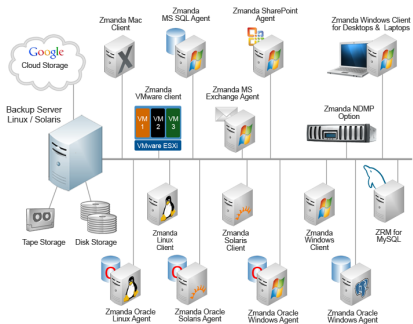


Figure: From Church, Harvard U.

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.
- ▶ Cost of storage often minor compared to cost of *access, processing, and data movement*.

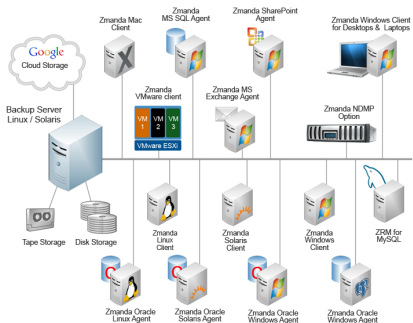


Figure: From Church, Harvard U.

The Era of Massive Data

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

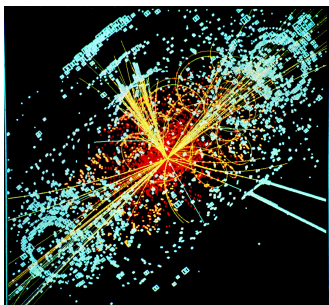


Figure: In search of the God particle, Wikipedia.

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.

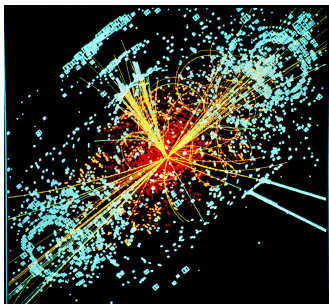


Figure: In search of the God particle, Wikipedia.

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.

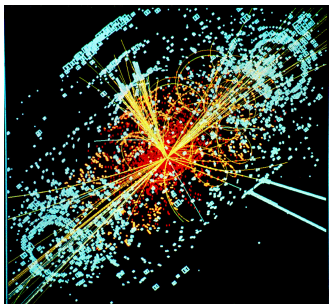


Figure: In search of the God particle, Wikipedia.

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...

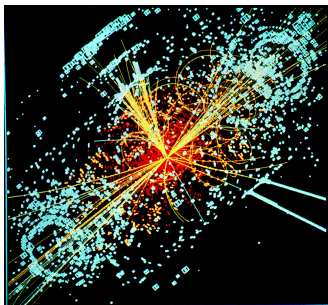


Figure: In search of the God particle, Wikipedia.

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

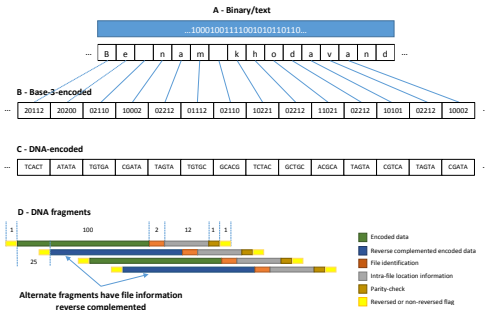
- ▶ **DNA is extremely durable:** Can still “read” mammoth, Neanderthal, 700,000 old horse DNA!
- ▶ **DNA write (synthesis) and read (sequencing) costs decrease daily.**
- ▶ **DNA information content of Human cell:** 6.4 GB. **Mass of a cell:** ~ 3 picograms. **No. of cells:** $15 - 40 \times 10^{12}$.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



Implementations

“Double Helix Serves Double Duty”, NY Times, Jan 2013

- Richard Feynman first to propose the use of macromolecules for storage (“There is plenty of room at the bottom”).



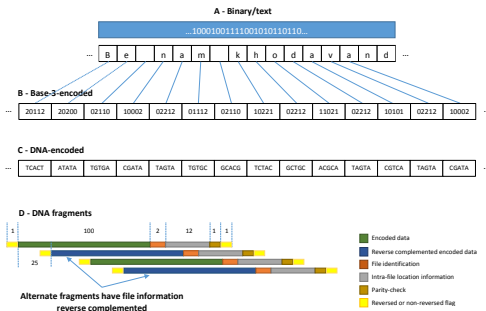
“Double Helix Serves Double Duty”, NY Times, Jan 2013

- 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.



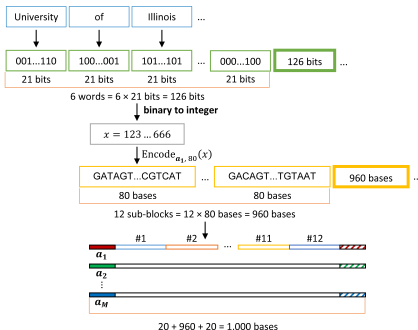
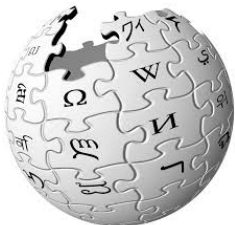
“Double Helix Serves Double Duty”, NY Times, Jan 2013

- 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.**



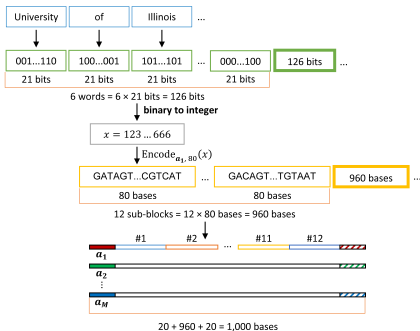
“Data Storage on DNA Can Keep it Safe for Centuries,” NY Times, Dec 2015

- 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.



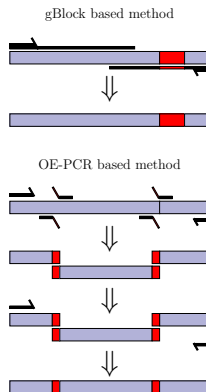
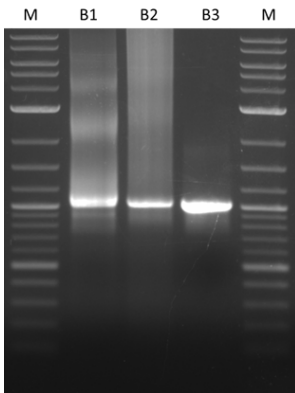
“Data Storage on DNA Can Keep it Safe for Centuries,” NY Times, Dec 2015

- 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).



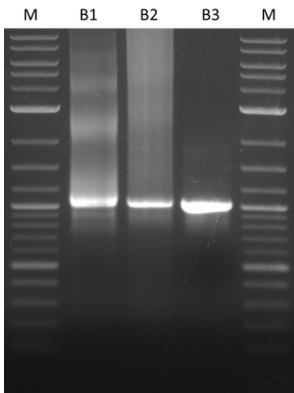
Our Experiments

- ▶ Random access achieved via **specialized address design**.

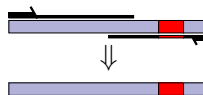


Our Experiments

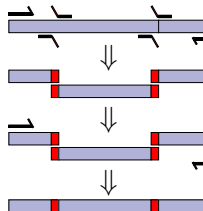
- ▶ Random access achieved via [specialized address design](#).
- ▶ Context identification and rewriting performed via [gBlock](#) or [OE-PCR](#) methods.



gBlock based method



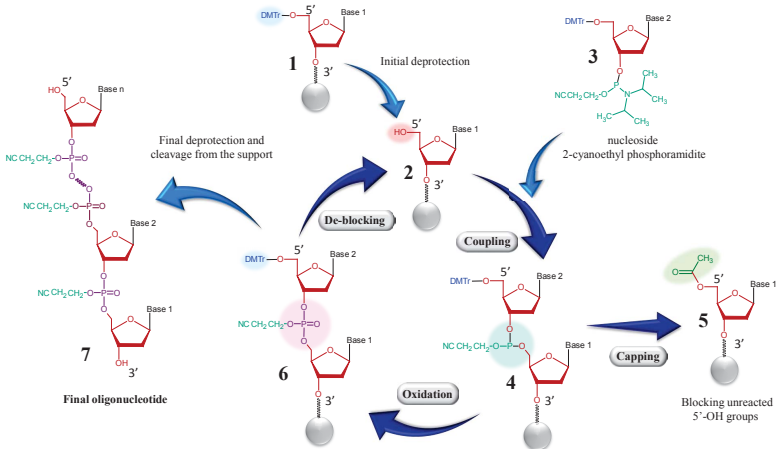
OE-PCR based method



The Write and Read Channels

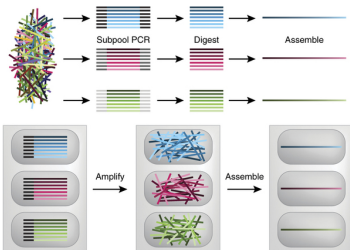
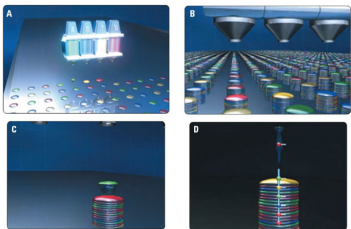
The Write Channel: DNA Synthesis

Biochemistry of synthesis: Adding bases through deprotection & coupling cycles.



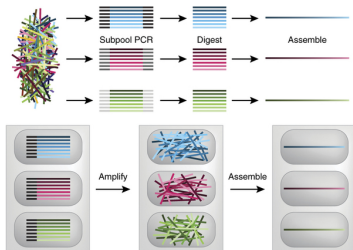
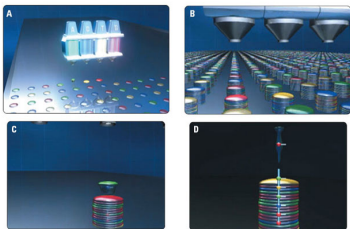
The Write Channel: DNA Synthesis

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



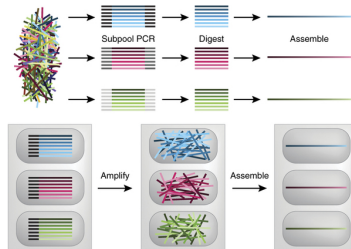
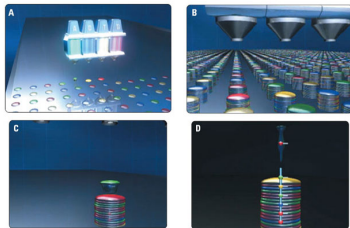
The Write Channel: DNA Synthesis

- ▶ **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.



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 Minlon

- ▶ **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.



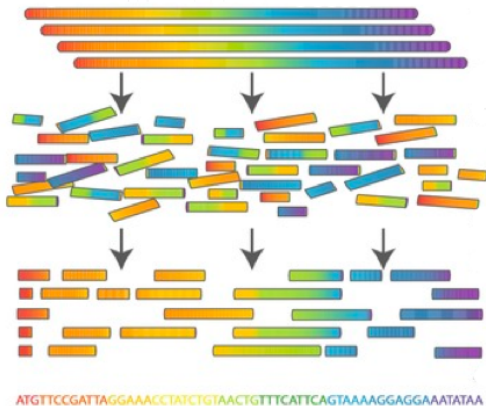
The Read Channel: Illumina and Minlon

- ▶ **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 - Minlon (Right):** Longer read length, miniaturized architecture. Large coverage errors, excessive number of block deletions.



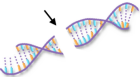
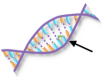
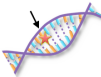
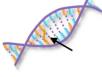
The Read Channel: Shotgun Sequencing

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



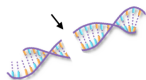
Media Aging

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

				
Type of Damage:	Double-strand break	Chemical bond between neighboring nucleotides	Chemical modification of a nucleotide	Chemical Linkage of Two Strands
Common Causes:	<ul style="list-style-type: none"> • Normal cellular activity • Ionizing radiation (including X-rays) • Chemotherapeutic drugs • DNA repair of other types of damage 	<ul style="list-style-type: none"> • Ultraviolet (UV) light 	<ul style="list-style-type: none"> • Reactive oxygen species (ROS) • Chemotherapeutic drugs • Other cellular and environmental chemicals • Normal modifications that regulate what genes are active 	<ul style="list-style-type: none"> • Reactive oxygen species (ROS) • Chemotherapeutic drugs • Other cellular and environmental chemicals

Media Aging

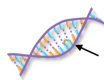
- ▶ **Breakage - Bursty Deletions - Transpositions/Reversals:** No built-in mechanism for correcting damages.
- ▶ **Coupled with synthesis and sequencing errors...**



Type of Damage: Double-strand break

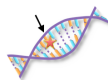
Common Causes:

- Normal cellular activity
- Ionizing radiation (including X-rays)
- Chemotherapeutic drugs
- DNA repair of other types of damage



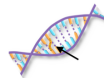
Chemical bond between neighboring nucleotides

- 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](#).

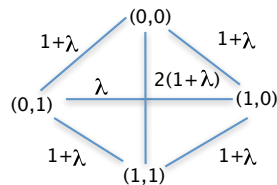
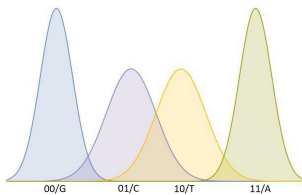
- ▶ A formal **mathematical theory of error-correction** for DNA storage?
 - ▶ Microarray Synthesis and Shotgun Sequencing: [DNA Profile Codes](#).
 - ▶ Microarray Synthesis and Nanopore Sequencing: [Asymmetric Lee Distance Codes](#).
 - ▶ DNA Media Aging: [Codes in the Damerau Distance](#).
- ▶ Mathematical 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](#).

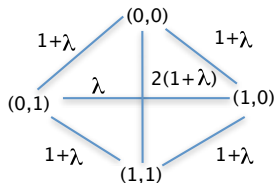
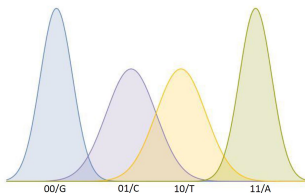
- ▶ A formal **mathematical theory of error-correction** for DNA storage?
 - ▶ Microarray Synthesis and Shotgun Sequencing: [DNA Profile Codes](#).
 - ▶ Microarray Synthesis and Nanopore Sequencing: [Asymmetric Lee Distance Codes](#).
 - ▶ DNA Media Aging: [Codes in the Damerau Distance](#).
- ▶ Mathematical 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 λ , the ALD $d_\lambda((\mathbf{a}; \mathbf{b}), (\mathbf{c}; \mathbf{d}))$ between pairs of binary sequences $(\mathbf{a}; \mathbf{b}), (\mathbf{c}; \mathbf{d})$ is defined as:

$$d_\lambda((\mathbf{a}; \mathbf{b}), (\mathbf{c}; \mathbf{d})) = \sum_{i=1}^n (1 + \lambda) (\mathbb{1}(a_i, b_i) + \mathbb{1}(c_i, d_i)) + \lambda \mathbb{1}(a_i, \bar{b}_i, \bar{c}_i, d_i) - 2(1 + \lambda) \mathbb{1}(a_i, b_i, c_i, d_i).$$

- ▶ 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.

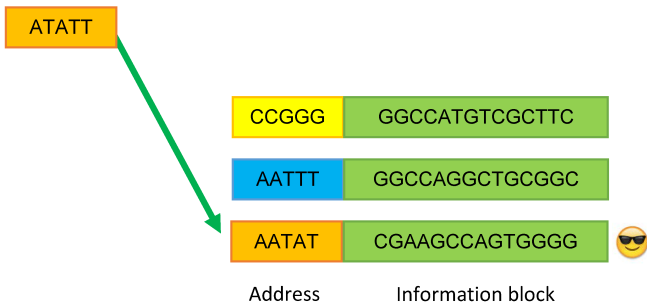
- ▶ A formal **mathematical theory of error-correction** for DNA storage?
 - ▶ DNA Media Aging: **Codes in the Damerau Distance**.
 - ▶ The **Damerau–Levenshtein (DL) distance** is a string metric, which for two strings over a finite alphabet equals the minimum number of insertions, deletions, substitutions 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-Tenengolt's extensions for the DL distance**: Uses the derivative of \mathbf{a} , $\mathbf{a}' = (a_1, a_2 + a_1, a_3 + a_2, \dots, a_n + a_{n-1})$.

- ▶ 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-Tenengolt's extensions for the DL distance**: Uses the derivative of \mathbf{a} , $\mathbf{a}' = (a_1, a_2 + a_1, a_3 + a_2, \dots, a_n + a_{n-1})$.
 - ▶ **Component codes**: $\mathcal{C}_H(n, 3)$ a single error-correcting code; $\mathcal{C}_D(n)$ a single deletion-correcting code.

$$\mathcal{C}_{T \vee D}(n) = \{\mathbf{a} \in \mathbb{F}_2^n : \mathbf{a} \in \mathcal{C}_D(n), \mathbf{a}' \in \mathcal{C}_H(n, 3)\}.$$

The code $\mathcal{C}_{T \vee D}(n)$ can correct one single deletion or adjacent transposition.

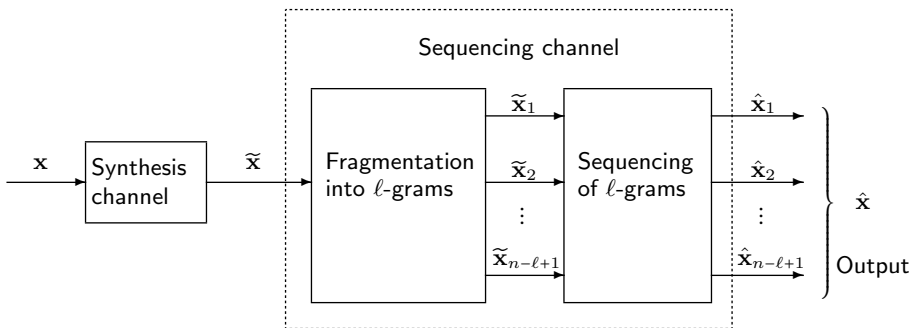
- ▶ Mathematical approaches for **enabling random access and rewriting?**
 - ▶ Address Design: **(Weakly) Mutually Uncorrelated Codes.**



- ▶ 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.

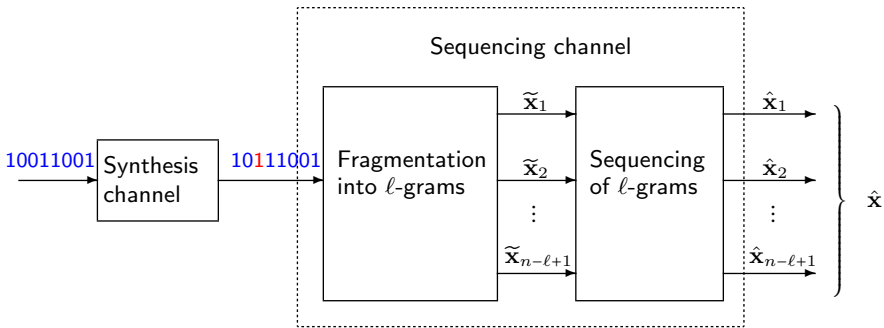
The DNA Storage Channel

DNA Storage Channel: Basics



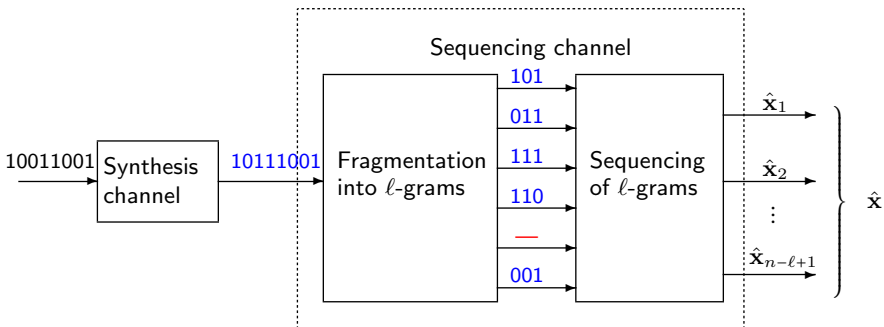
Synthesis channel captures the “write” process.

DNA Storage Channel: Basics



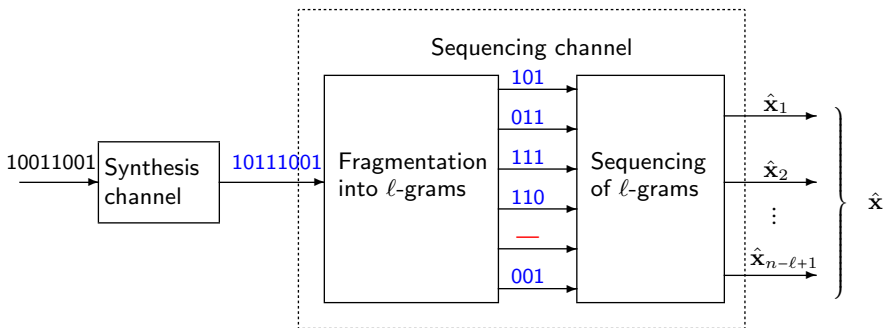
The sequence synthesis process introduces errors (current technologies $\leq 0.1\%$).

DNA Storage Channel: Basics



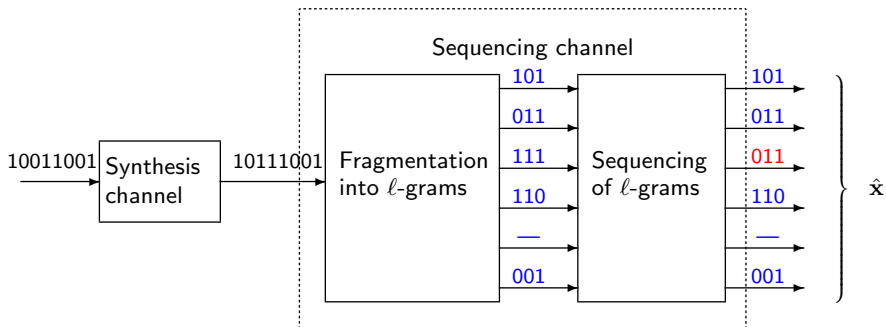
DNA sequencing represents the “read” process. Consists of fragmenting sequence to be read, and “reconstructing” fragments (ℓ -grams).

DNA Storage Channel: Basics



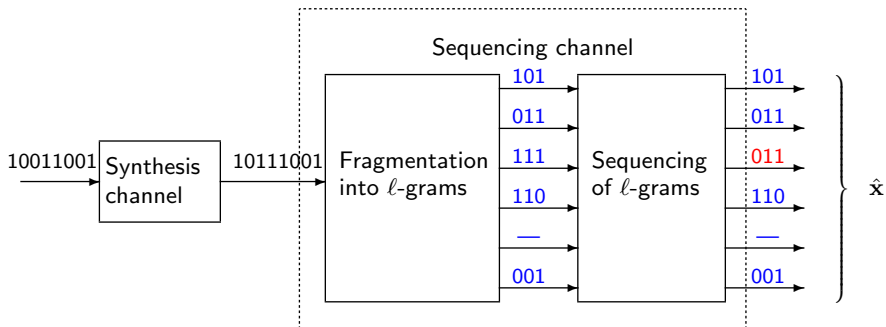
Note that strings at the output of the fragmentation block are **not** ordered.

DNA Storage Channel: Basics



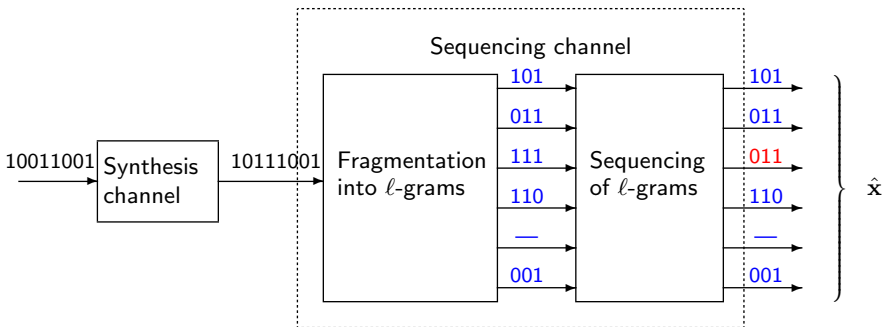
- Sequencing introduces **errors in some l -grams** and **some l -grams may not be covered**.

DNA Storage Channel: Basics



- ▶ Sequencing introduces **errors in some l -grams** and **some l -grams may not be covered**.
- ▶ Modern Illumina platforms have substitution error rates $\leq 0.5\%$. Coverage errors context-dependent.

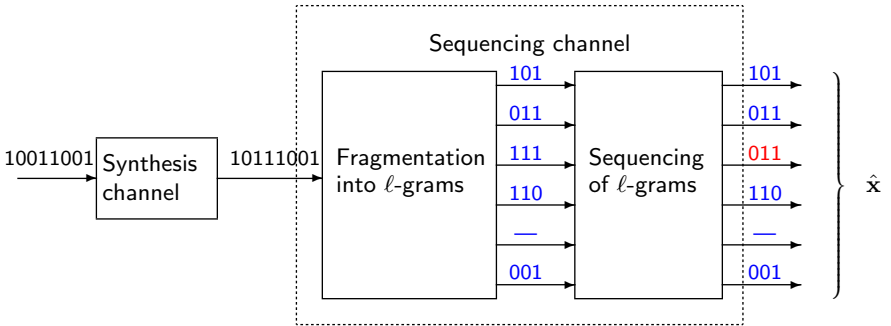
DNA Storage Channel: Profile Vectors



Profile vectors

The profile vector of a sequence reflects the **count of its l -grams**;

DNA Storage Channel: Profile Vectors



Profile vectors

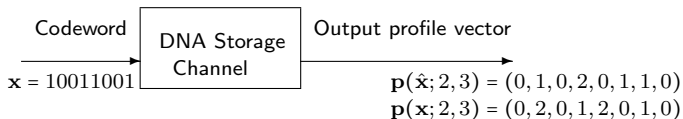
The profile vector of a sequence reflects the **count of its l -grams**;

Example

Profile of vector $x = 10011001$ equals

000	001	010	011	100	101	110	111
(0,	2,	0,	1,	2,	0,	1,	0).

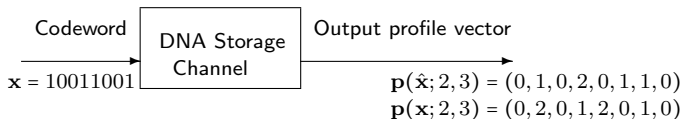
Input and Output Profile Vectors



Profile Vectors: Formal Definition

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

Input and Output Profile Vectors



Profile Vectors: Formal Definition

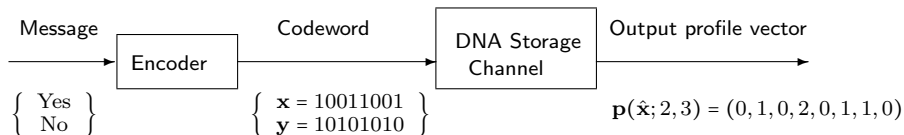
Fix **alphabet size** q and fragment (read) length $\ell < n$. The **profile vector** of some sequence \mathbf{x} , denoted by $\mathbf{p}(\mathbf{x}; q, \ell)$, has length q^ℓ and its entry indexed by \mathbf{z} equals the number of occurrences of \mathbf{z} in \mathbf{x} as an ℓ -gram.

Example

Profile of $\mathbf{x} = 10011001$ and sequencing channel output:

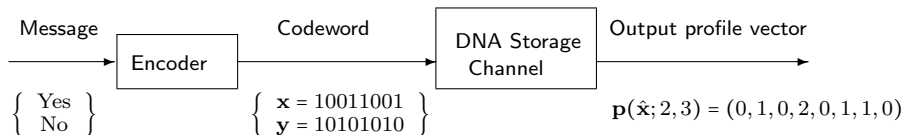
000	001	010	011	100	101	110	111
(0,	2,	0,	1,	2,	0,	1,	0),
(0,	1,	0,	2,	0,	1,	1,	0).

Code Design Criteria



		000	001	010	011	100	101	110	111
$\mathbf{p}(\mathbf{x}; 2, 3)$	=	(0,	2,	0,	1,	2,	0,	1,	0)
$\mathbf{p}(\mathbf{y}; 2, 3)$	=	(0,	0,	3,	0,	0,	3,	0,	0)
$\mathbf{p}(\hat{\mathbf{x}}; 2, 3)$	=	(0,	1,	0,	2,	0,	1,	1,	0).

Code Design Criteria

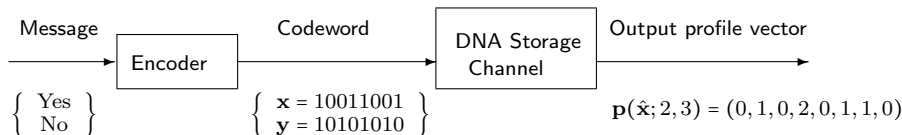


$$\begin{array}{rcl}
 \mathbf{p}(\mathbf{x}; 2, 3) & = & \begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 2, & 0, & 1, & 2, & 0, & 1, & 0) \end{array} \\
 \mathbf{p}(\mathbf{y}; 2, 3) & = & \begin{array}{cccccccc} (0, & 0, & 3, & 0, & 0, & 3, & 0, & 0) \end{array} \\
 \mathbf{p}(\hat{\mathbf{x}}; 2, 3) & = & \begin{array}{cccccccc} (0, & 1, & 0, & 2, & 0, & 1, & 1, & 0). \end{array}
 \end{array}$$

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 ℓ -gram errors**.

Code Design Criteria



$$\begin{array}{rcl}
 \mathbf{p}(\mathbf{x}; 2, 3) & = & \begin{array}{cccccccc} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 2, & 0, & 1, & 2, & 0, & 1, & 0) \end{array} \\
 \mathbf{p}(\mathbf{y}; 2, 3) & = & \begin{array}{cccccccc} (0, & 0, & 3, & 0, & 0, & 3, & 0, & 0) \end{array} \\
 \mathbf{p}(\hat{\mathbf{x}}; 2, 3) & = & \begin{array}{cccccccc} (0, & 1, & 0, & 2, & 0, & 1, & 1, & 0). \end{array}
 \end{array}$$

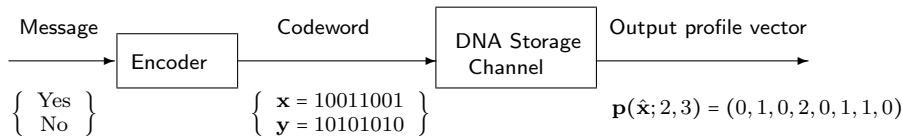
Condition 1

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

Definition 1

The **ℓ -gram distance** between \mathbf{x} and \mathbf{y} equals the **asymmetric distance** (ℓ_1 distance) between $\mathbf{p}(\mathbf{x}; q, \ell)$ and $\mathbf{p}(\mathbf{y}; q, \ell)$.

Code Design Criteria

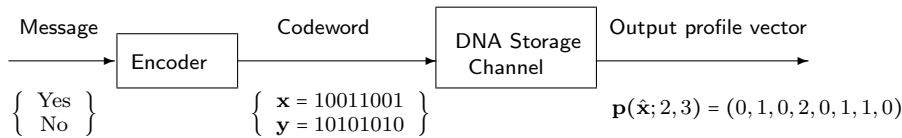


		000	001	010	011	100	101	110	111
$\mathbf{p}(\mathbf{x}; 2, 3)$	=	(0,	2,	0,	1,	2,	0,	1,	0)
$\mathbf{p}(\mathbf{y}; 2, 3)$	=	(0,	0,	3,	0,	0,	3,	0,	0)
$\mathbf{p}(\hat{\mathbf{x}}; 2, 3)$	=	(0,	1,	0,	2,	0,	1,	1,	0).

Definition 1

The ℓ -gram distance between \mathbf{x} and \mathbf{y} equals the asymmetric distance (ℓ_1 distance) between $\mathbf{p}(\mathbf{x}; q, \ell)$ and $\mathbf{p}(\mathbf{y}; q, \ell)$.

Code Design Criteria



$$\begin{array}{rcl}
 \mathbf{p}(\mathbf{x}; 2, 3) & = & \begin{array}{cccccccc} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 2, & 0, & 1, & 2, & 0, & 1, & 0) \end{array} \\
 \mathbf{p}(\mathbf{y}; 2, 3) & = & \begin{array}{cccccccc} (0, & 0, & 3, & 0, & 0, & 3, & 0, & 0) \end{array} \\
 \mathbf{p}(\hat{\mathbf{x}}; 2, 3) & = & \begin{array}{cccccccc} (0, & 1, & 0, & 2, & 0, & 1, & 1, & 0). \end{array}
 \end{array}$$

Definition 1

The ℓ -gram distance between \mathbf{x} and \mathbf{y} equals the asymmetric distance (ℓ_1 distance) between $\mathbf{p}(\mathbf{x}; q, \ell)$ and $\mathbf{p}(\mathbf{y}; q, \ell)$.

Asymmetric Distance

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

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

Code Design Criteria

$$\begin{array}{rcl}
 & & 000 \quad 001 \quad 010 \quad 011 \quad 100 \quad 101 \quad 110 \quad 111 \\
 \mathbf{p}(\mathbf{x}; 2, 3) & = & (0, \quad 2, \quad 0, \quad 1, \quad 2, \quad 0, \quad 1, \quad 0) \\
 \mathbf{p}(\mathbf{y}; 2, 3) & = & (0, \quad 0, \quad 3, \quad 0, \quad 0, \quad 3, \quad 0, \quad 0) \\
 \mathbf{p}(\hat{\mathbf{x}}; 2, 3) & = & (0, \quad 1, \quad 0, \quad 2, \quad 0, \quad 1, \quad 1, \quad 0).
 \end{array}$$

Asymmetric Distance

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

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

Code Design Criteria

$$\begin{array}{rcl}
 & & 000 \quad 001 \quad 010 \quad 011 \quad 100 \quad 101 \quad 110 \quad 111 \\
 \mathbf{p}(\mathbf{x}; 2, 3) & = & (0, \quad 2, \quad 0, \quad 1, \quad 2, \quad 0, \quad 1, \quad 0) \\
 \mathbf{p}(\mathbf{y}; 2, 3) & = & (0, \quad 0, \quad 3, \quad 0, \quad 0, \quad 3, \quad 0, \quad 0) \\
 \mathbf{p}(\hat{\mathbf{x}}; 2, 3) & = & (0, \quad 1, \quad 0, \quad 2, \quad 0, \quad 1, \quad 1, \quad 0).
 \end{array}$$

Asymmetric Distance

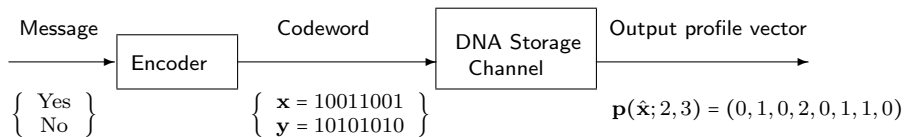
Let $\mathbf{u}, \mathbf{v} \in \mathbb{Z}_{\geq 0}^N$. Define $\Delta(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^N \max(u_i - v_i, 0)$.

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

Minimum Asymmetric Distance

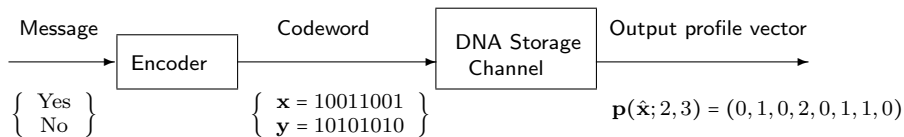
A DNA storage code with minimum asymmetric distance d can correct s_1 substitution errors due to synthesis, s_2 substitution errors due to sequencing and t coverage errors provided that $d > 2s_1 + 2s_2 + t$.

Code Design Criteria



		000	001	010	011	100	101	110	111
$\mathbf{p}(\mathbf{x}; 2, 3)$	=	(0,	2,	0,	1,	2,	0,	1,	0)
$\mathbf{p}(\mathbf{y}; 2, 3)$	=	(0,	0,	3,	0,	0,	3,	0,	0)

Code Design Criteria

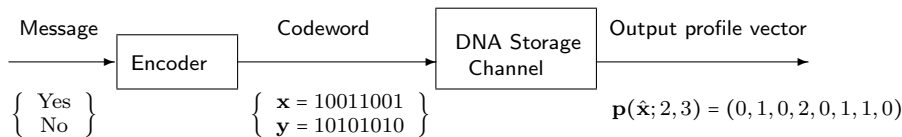


$$\begin{array}{rcl}
 \mathbf{p}(\mathbf{x}; 2, 3) & = & \begin{array}{cccccccc} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 2, & 0, & 1, & 2, & 0, & 1, & 0) \end{array} \\
 \mathbf{p}(\mathbf{y}; 2, 3) & = & \begin{array}{cccccccc} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 0, & 3, & 0, & 0, & 3, & 0, & 0) \end{array}
 \end{array}$$

Condition 2

Codewords whose ℓ -grams avoid error-causing substrings.

Code Design Criteria



$$\begin{array}{rcl}
 \mathbf{p}(\mathbf{x}; 2, 3) & = & \begin{array}{cccccccc} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 2, & 0, & 1, & 2, & 0, & 1, & 0) \end{array} \\
 \mathbf{p}(\mathbf{y}; 2, 3) & = & \begin{array}{cccccccc} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 0, & 3, & 0, & 0, & 3, & 0, & 0) \end{array}
 \end{array}$$

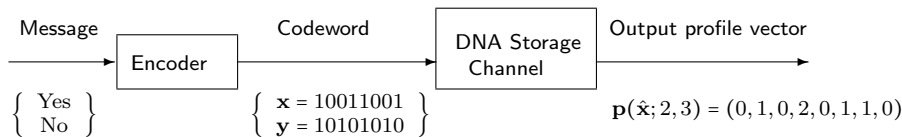
Condition 2

Codewords whose ℓ -grams avoid error-causing substrings.

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

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

Code Design Criteria



$$\begin{array}{rcl}
 \mathbf{p}(\mathbf{x}; 2, 3) & = & \begin{array}{cccccccc} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 2, & 0, & 1, & 2, & 0, & 1, & 0) \end{array} \\
 \mathbf{p}(\mathbf{y}; 2, 3) & = & \begin{array}{cccccccc} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 0, & 3, & 0, & 0, & 3, & 0, & 0) \end{array}
 \end{array}$$

Condition 2

Codewords whose ℓ -grams avoid error-causing substrings.

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

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

For example, may require that ℓ -grams lie in

$$S = \{001, 010, 011, 100, 101, 110\}.$$

Fundamental Questions

Distinct ℓ -gram Profile Vectors

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

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

Fundamental Questions

ℓ -gram Reconstruction Code (GRC)

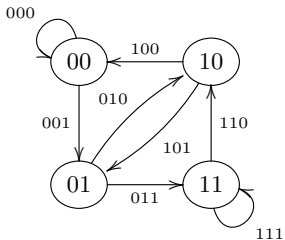
$\mathcal{C} \subseteq \mathcal{Q}(n; S)$ is an $(n, d; S)$ - ℓ -GRC if the ℓ -gram distance between any pair of distinct words is at least d .

Construct good $(n, d; S)$ - ℓ -GRC. “Good” means large codebook size, avoidance of bad ℓ -grams.

Profile Vectors and ℓ -Gram Codes

De Bruijn Graphs

Example for $q = 2$, $\ell = 3$.



De Bruijn Graphs

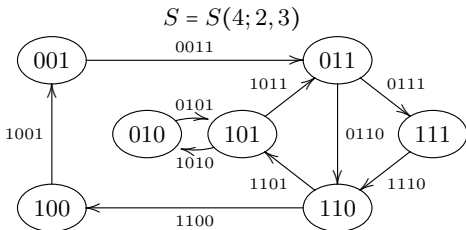
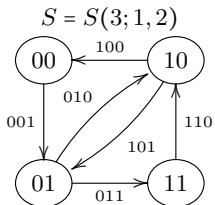
Nodes are q -ary strings of length $\ell - 1$.

$(\mathbf{v}, \mathbf{v}')$ is an arc if

$$\begin{array}{ccccccc}
 v_2 & v_3 & & & v_{\ell-1} & & \\
 \parallel & \parallel & \dots & & \parallel & & \\
 v'_1 & v'_2 & & & v'_{\ell-2} & &
 \end{array}$$

Restricted De Bruijn Graphs (Ruskey *et al.*, 2012)

Let $S(\ell; w_1, w_2)$ denote the binary strings of length ℓ with weight between w_1 and w_2 .



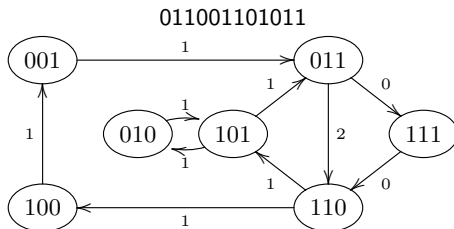
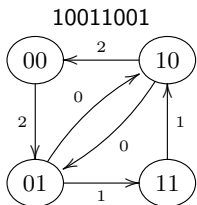
Restricted De Bruijn Graphs $D(S)$

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

$(\mathbf{v}, \mathbf{v}')$ is an arc if

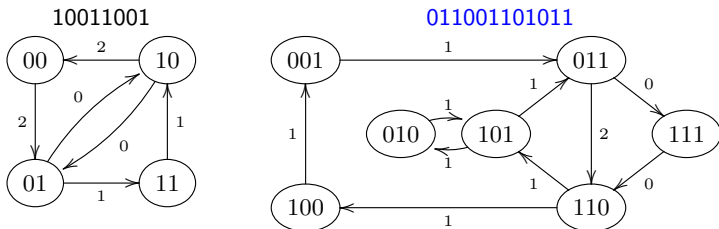
$$\begin{array}{ccccccc}
 v_2 & v_3 & & v_{\ell-1} & & & \\
 \parallel & \parallel & \dots & \parallel & \text{and} & v_1 v_2 \dots v_{\ell-1} v'_{\ell-1} \in S. & \\
 v'_1 & v'_2 & & v'_{\ell-2} & & &
 \end{array}$$

Profile Vectors and Flows



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

Profile Vectors and Flows



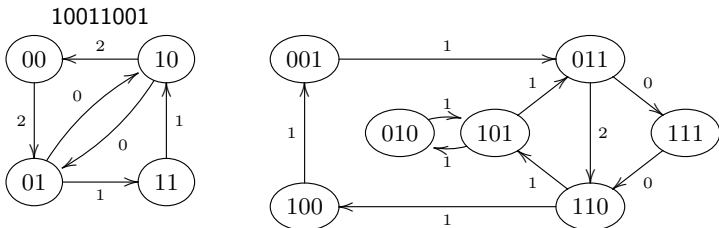
Representation of profile vectors of words in $\mathcal{Q}(n; S)$ using the digraph $D(S)$.

Closed Words

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

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

Profile Vectors and Flows



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

Closed Words

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

$\overline{\mathcal{Q}}(n; S)$: largest set of q -ary closed words of length n whose ℓ -grams belong to S and which have distinct ℓ -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 $\overline{\mathcal{Q}}(n; S)$ 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.

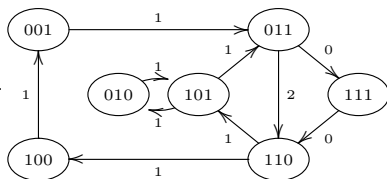
Flow conservations equations:

$$\mathbf{B}\mathbf{u} = \mathbf{0},$$

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

Sum of flows:

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



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

$$\mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b} \text{ and } \mathbf{u} \geq \mathbf{0}.$$

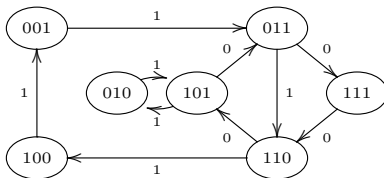
Sufficient Conditions

Flows are not always profile vectors

Let $\mathbf{u} \geq \mathbf{0}$ be such that

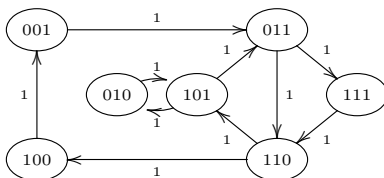
$$\mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b}.$$

This does **not** imply that \mathbf{u} is a profile vector!



Sufficient Conditions

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



Profile vector of **0110101110011**.

$$\mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b} \text{ and } \mathbf{u} > \mathbf{0}.$$

Profile Vectors and Lattice Points

Consider the following two sets of lattice points:

$$\mathcal{F}(n; S) = \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b}, \mathbf{u} \geq \mathbf{0}\},$$

$$\mathcal{E}(n; S) = \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b}, \mathbf{u} > \mathbf{0}\}.$$

Clearly, one has

$$|\mathcal{E}(n; S)| \leq |\overline{\mathcal{Q}}(n; S)| \leq |\mathcal{F}(n; S)|.$$

Profile Vectors and Lattice Points

Consider the following two sets of lattice points:

$$\mathcal{F}(n; S) = \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b}, \mathbf{u} \geq \mathbf{0}\},$$

$$\mathcal{E}(n; S) = \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b}, \mathbf{u} > \mathbf{0}\}.$$

Clearly, one has

$$|\mathcal{E}(n; S)| \leq |\overline{\mathcal{Q}}(n; S)| \leq |\mathcal{F}(n; S)|.$$

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 $|\mathcal{E}(n; S)|, |\mathcal{F}(n; S)|$.

Profile Vectors and Lattice Points

Consider the following two sets of lattice points:

$$\mathcal{F}(n; S) \triangleq \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b}, \mathbf{u} \geq \mathbf{0}\},$$

$$\mathcal{E}(n; S) \triangleq \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b}, \mathbf{u} > \mathbf{0}\}.$$

$$|\mathcal{E}(n; S)| \leq |\overline{\mathcal{Q}}(n; S)| \leq |\mathcal{F}(n; S)|.$$

Profile Vectors and Lattice Points

Consider the following two sets of lattice points:

$$\mathcal{F}(n; S) \triangleq \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b}, \mathbf{u} \geq \mathbf{0}\},$$

$$\mathcal{E}(n; S) \triangleq \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b}, \mathbf{u} > \mathbf{0}\}.$$

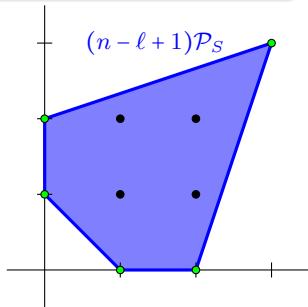
$$|\mathcal{E}(n; S)| \leq |\overline{\mathcal{Q}}(n; S)| \leq |\mathcal{F}(n; S)|.$$

Observations

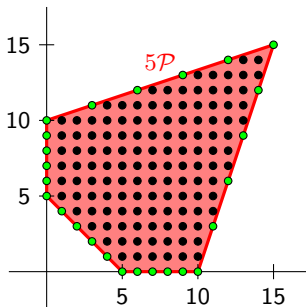
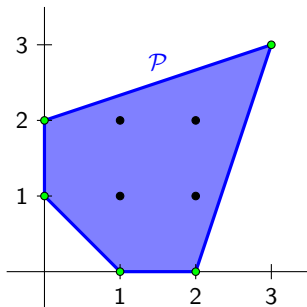
- Define the **polytope**

$$\mathcal{P}_S = \{\mathbf{u} \in \mathbb{R}^{|S|} : \mathbf{A}\mathbf{u} = \mathbf{b}, \mathbf{u} \geq \mathbf{0}\}.$$

- $\mathcal{F}(n; S)$ is the set of lattice points in $(n - \ell + 1)\mathcal{P}_S$.
- $\mathcal{E}(n; S)$ is the set of lattice points in the **interior** of $(n - \ell + 1)\mathcal{P}_S$.



Lattice Point Enumeration in Dilated Polytopes



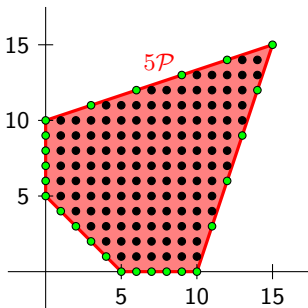
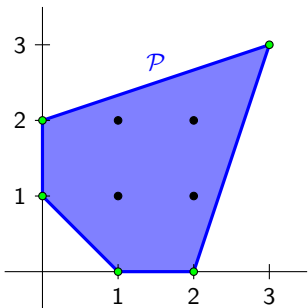
For a polytope $\mathcal{P} \subset \mathbb{R}^N$ and $t \in \mathbb{R}$, the **dilation** $t\mathcal{P}$ is given by

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

The **lattice point enumerator** for \mathcal{P} is $\mathcal{L}_{\mathcal{P}} : \mathbb{R} \rightarrow \mathbb{Z}$ 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

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

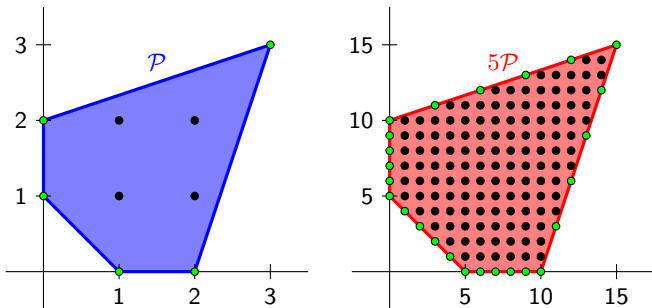
The **lattice point enumerator** for \mathcal{P} is $\mathcal{L}_{\mathcal{P}} : \mathbb{R} \rightarrow \mathbb{Z}$ defined by

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

Theorem (Ehrhart)

If \mathcal{P} is a rational D -dimensional polytope, then $\mathcal{L}_{\mathcal{P}}(t)$ is a “quasipolynomial” (polynomial with periodic functions as coefficients) in t of degree D .

Lattice Point Enumeration in Dilated Polytopes



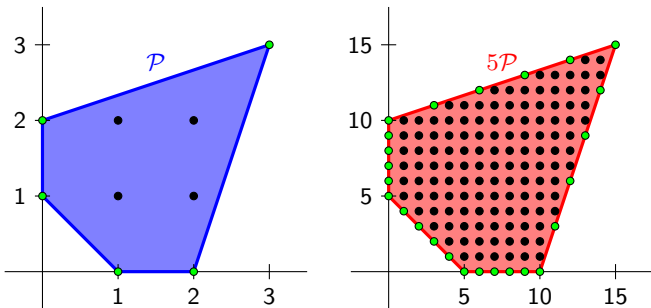
For a polytope $\mathcal{P} \subset \mathbb{R}^N$ and $t \in \mathbb{R}$, the **dilation** $t\mathcal{P}$ is given by

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

The **lattice point enumerator** for \mathcal{P} is $\mathcal{L}_{\mathcal{P}} : \mathbb{R} \rightarrow \mathbb{Z}$ 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 $\mathcal{L}_{\mathcal{P}} : \mathbb{R} \rightarrow \mathbb{Z}$ 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 .

Main Enumeration Results

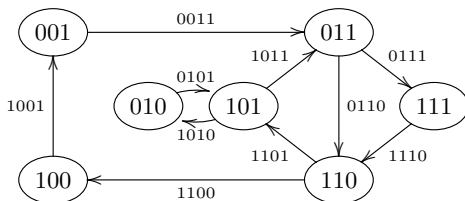
Theorem

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, $|\overline{\mathcal{Q}}(n; S)| = \Theta'(n^{|S| - |V(S)|})$.

Main Enumeration Results

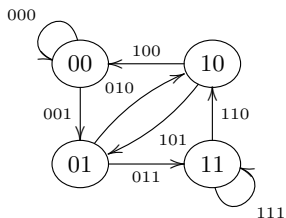
Theorem

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, $|\overline{\mathcal{Q}}(n; S)| = \Theta'(n^{|S| - |V(S)|})$.



Here, $|\overline{\mathcal{Q}}(n; S)| = \Theta'(n^3)$.

Corollaries of Main Enumeration Result



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

Theorem (Jacquet, Knessl, Szpankowski, 2012; Ukkonen, Pevzner 1990's)

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

$$|\mathcal{E}(n; S)| \sim |\mathcal{F}(n; S)| \sim |\overline{\mathcal{Q}}(n; S)| \sim c(S)n^{q^\ell - q^{\ell-1}} \text{ where } c(S) \text{ is a constant.}$$

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

Corollary

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

$$|\mathcal{E}(n; S)| \sim |\mathcal{F}(n; S)| \sim |\overline{\mathcal{Q}}(n; S)| \sim c(S)n^{|S| - |V|} \text{ where } c(S) \text{ is a constant.}$$

Varshamov Codes

Fix d and let p be a prime such that $p > d$ and $p > N$. Choose N distinct nonzero elements $\alpha_1, \alpha_2, \dots, \alpha_N$ in $\mathbb{Z}/p\mathbb{Z}$ and consider the matrix

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

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

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

Varshamov Codes

Fix d and let p be a prime such that $p > d$ and $p > N$. Choose N distinct nonzero elements $\alpha_1, \alpha_2, \dots, \alpha_N$ in $\mathbb{Z}/p\mathbb{Z}$ and consider the matrix

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

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

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

Theorem (Varshamov, 1973)

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

Gram Reconstruction Codes

Construction I

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

Gram Reconstruction Codes

Construction I

Let $\mathbf{pQ}(n; S)$ be the set of distinct profile vectors of words in S and $N = |S|$. Then $\mathcal{C}(\mathbf{H}, \boldsymbol{\beta}) \cap \mathbf{pQ}(n; S)$ is an $(n, d + 1; S)$ - ℓ -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

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

Then $\mathcal{C}(\mathbf{H}, \boldsymbol{\beta})$ contains the following words.

$$\begin{array}{llll} (4, 0, 0, 1, 0, 1) & & (0, 1, 1, 4, 0, 0) & \\ (2, 2, 0, 2, 0, 0) & \leftrightarrow & 00100100 & (0, 1, 0, 0, 4, 1) \\ (1, 4, 0, 0, 1, 0) & & (0, 0, 4, 1, 1, 0) & \\ (1, 1, 1, 1, 1, 1) & \leftrightarrow & 00101100 & (0, 0, 2, 0, 2, 2) \leftrightarrow 01101101 \\ (1, 0, 1, 0, 0, 4) & & & \end{array}$$

Above, **three** profile vectors in $\mathbf{pQ}(8; S)$.

Gram Reconstruction Codes

Construction I

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

Gram Reconstruction Codes

Construction I

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

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

However, the **optimal choice of β is not known.**

Gram Reconstruction Codes

Construction I

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

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

However, the **optimal choice of β is not known**.

We fix a certain choice of \mathbf{H} and β and provide lower bounds on the size of $\mathcal{C}(\mathbf{H}, \beta) \cap \mathbf{pQ}(n; S)$ as a function of n .

Ehrhart Theory Continued

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

$$\mathbf{A}_{\text{GRC}} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{H} & -p\mathbf{I}_d \end{array} \right).$$

Proposition

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

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

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

Ehrhart Theory Continued

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

$$\mathbf{A}_{\text{GRC}} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{H} & -p\mathbf{I}_d \end{array} \right).$$

Proposition

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

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

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

Theorem

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

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

- ▶ $f(n) = \Omega'(g(n))$ means that for a fixed value of ℓ , there exists an integer λ 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).

Literature Overview

- 1 H. M. Kiah, G. J. Puleo, and O. Milenkovic. [Codes for DNA Sequence Profiles](#), IEEE Trans. Info. Theory (2016)
- 2 S. M. H. Tabatabaei Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic. [A Rewritable, Random-Access DNA-Based Storage System](#), Nature SR (2015)
- 3 R. Gabrys, H. M. Kiah, and O. Milenkovic. [Asymmetric Lee Distance Codes for DNA-Based Storage](#), CoRR abs/1506.00740 (2015)
- 4 S. M. Hossein Tabatabaei Yazdi, H. M. Kiah, E. R. Garcia, J. Ma, H. Zhao, and O. Milenkovic. [DNA-Based Storage: Trends and Methods](#), IEEE Trans. on Molecular Communication (2016)
- 5 R. Gabrys, E. Yaakobi, and O. Milenkovic. [Codes in the Damerau Distance for DNA-Based Storage](#), preprint.
- 6 S. M. H. Tabatabaei Yazdi, H. M. Kiah and O. Milenkovic, [Weakly Mutually Uncorrelated Codes](#), preprint.

THANK YOU!