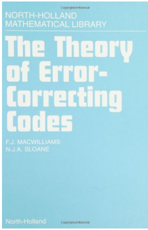
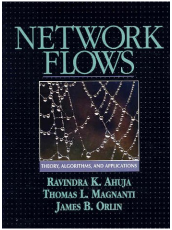
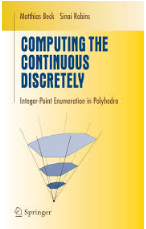
Part 3: Coding for DNA-Based Data Storage

Olgica Milenkovic
University of Illinois, Urbana-Champaign
North American School of Information Theory, Texas, 2018

May 2018







Coding Problems

Fundamentally New Coding Questions

- DNA Profile and Uniquely Reconstructable Codes (IDT Synthesis, Illumina Sequencing)

  Codes for DNA Sequence Profiles, IT, 2016; Unique Reconstruction of Coded Strings from Multiset Substring Spectra, ISIT 2018.

Fundamentally New Coding Questions

- DNA Profile and Uniquely Reconstructable Codes (IDT Synthesis, Illumina Sequencing)

  Codes for DNA Sequence Profiles, IT, 2016; Unique Reconstruction of Coded Strings from Multiset Substring Spectra, ISIT 2018.

- Address Design for Random Access (All Platforms)

  Mutually Uncorrelated Primers for DNA-Based Data Storage, IT 2018.

Fundamentally New Coding Questions

- **DNA Profile and Uniquely Reconstructable Codes (IDT Synthesis, Illumina Sequencing)**

  Codes for DNA Sequence Profiles, IT, 2016; Unique Reconstruction of Coded Strings from Multiset Substring Spectra, ISIT 2018.

- **Address Design for Random Access (All Platforms)**

  Mutually Uncorrelated Primers for DNA-Based Data Storage, IT 2018.

- **Coding for Nanopore readout systems (MinION and Solid State)**

  Asymmetric Lee Distance Codes for DNA-Based Storage, IT 2017;
  The Hybrid k-Deck Problem: Reconstructing Sequences from Short and Long Traces, ISIT 2017.

Fundamentally New Coding Questions

- **DNA Profile and Uniquely Reconstructable Codes (IDT Synthesis, Illumina Sequencing)**

  Codes for DNA Sequence Profiles, IT, 2016; Unique Reconstruction of Coded Strings from Multiset Substring Spectra, ISIT 2018.

- **Address Design for Random Access (All Platforms)**

  Mutually Uncorrelated Primers for DNA-Based Data Storage, IT 2018.

- **Coding for Nanopore readout systems (MinION and Solid State)**

  Asymmetric Lee Distance Codes for DNA-Based Storage, IT 2017;
  The Hybrid k-Deck Problem: Reconstructing Sequences from Short and Long Traces, ISIT 2017.

- **Small-Intersection Set Discrepancy (Nicking)**

  Manuscript in preparation, 2018.

Fundamentally New Coding Questions

- **DNA Profile and Uniquely Reconstructable Codes (IDT Synthesis, Illumina Sequencing)**

  Codes for DNA Sequence Profiles, IT, 2016; Unique Reconstruction of Coded Strings from Multiset Substring Spectra, ISIT 2018.

- **Address Design for Random Access (All Platforms)**

  Mutually Uncorrelated Primers for DNA-Based Data Storage, IT 2018.

- **Coding for Nanopore readout systems (MinION and Solid State)**

  Asymmetric Lee Distance Codes for DNA-Based Storage, IT 2017;
  The Hybrid k-Deck Problem: Reconstructing Sequences from Short and Long Traces, ISIT 2017.

- **Small-Intersection Set Discrepancy (Nicking)**

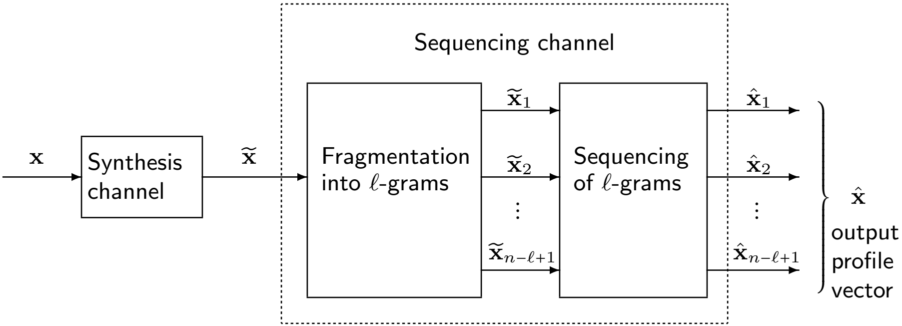
  Manuscript in preparation, 2018.

- **Codes in the Damerau Distance (Aging)**

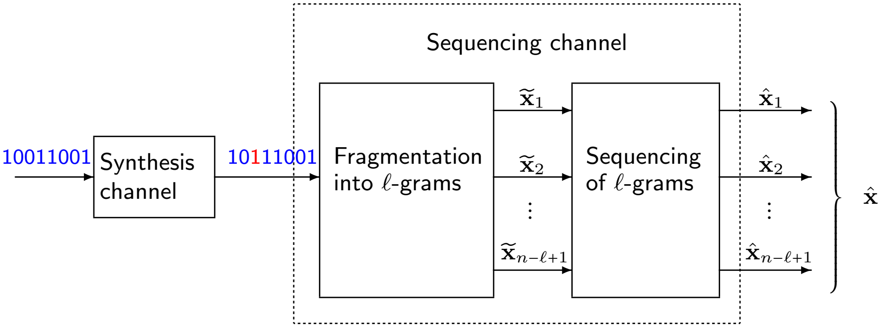
  Codes in the Damerau Distance for Deletion and Adjacent Transposition Correction, IT 2018.

DNA Profile Codes

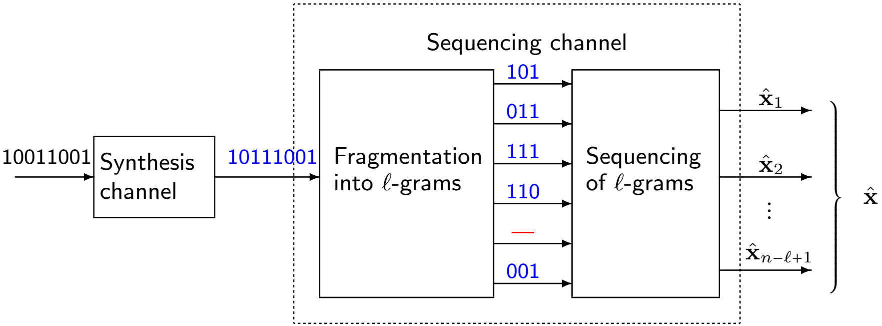
DNA Storage Channel



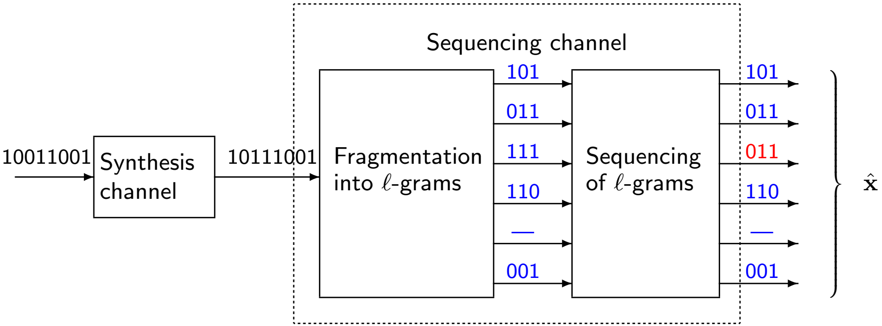
DNA Storage Channel



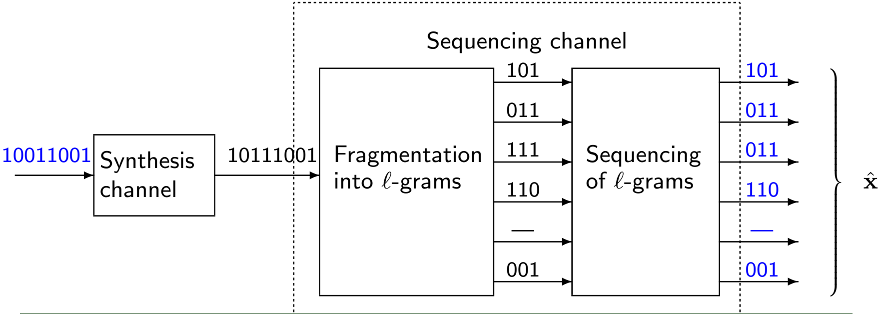
DNA Storage Channel



DNA Storage Channel



DNA Storage Channel – Output Profile Vectors



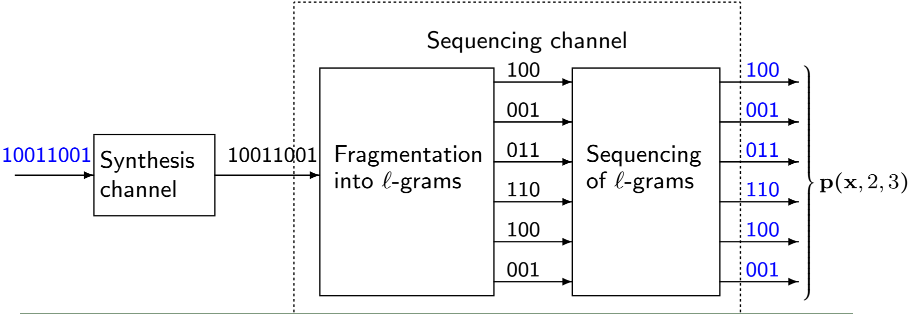
Output profile vector

Given an input sequence 10011001, we obtain an output profile vector that reflects the (possibly erroneous) count of each substring

$$\begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 1, & 0, & 2, & 0, & 1, & 1, & 0). \end{array}$$

Note: position of substring is not known!
Note: input is a binary sequence, while the output is an integer-valued vector.

DNA Storage Channel – Profile Vectors

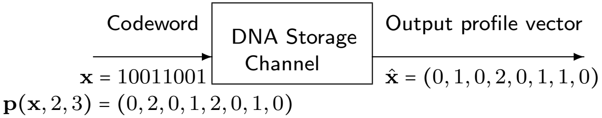


Profile vector

Given an input sequence $\mathbf{x}$ = 10011001, its profile vector denoted by $\mathbf{p}(\mathbf{x}, q, \ell)$ reflects the actual count of each substring

$$\begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 2, & 0, & 1, & 2, & 0, & 1, & 0). \end{array}$$

DNA Storage Channel – Profile Vectors

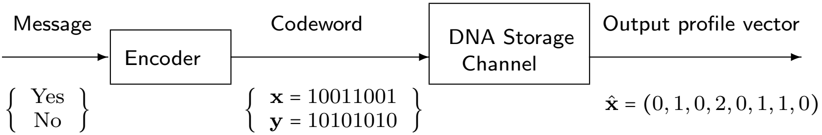


Profile vector

Given an input sequence $\mathbf{x} = 10011001$, its profile vector denoted by $\mathbf{p}(\mathbf{x}, q, \ell)$ reflects the actual count of each substring

$$\begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ (0, & 2, & 0, & 1, & 2, & 0, & 1, & 0). \end{array}$$

Code Design Criteria



$$
\begin{array}{lcccccccccl}
 & & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 & \\
\mathbf{p}(\mathbf{x};2,3) & = & (0, & 2, & 0, & 1, & 2, & 0, & 1, & 0) & \leftarrow \text{dist with } \hat{\mathbf{x}} \text{ is } 3 \\
\mathbf{p}(\mathbf{y};2,3) & = & (0, & 0, & 3, & 0, & 0, & 3, & 0, & 0) & \leftarrow \text{dist with } \hat{\mathbf{x}} \text{ is } 5
\end{array}
$$

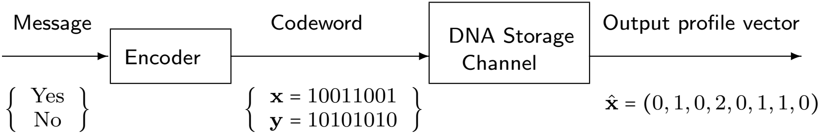
Criterion 1

Codewords whose profile vectors are far from each other.

We define the ℓ-gram distance between $\mathbf{x}$ and $\mathbf{y}$ to be the asymmetric distance between the profile vectors of $\mathbf{x}$ and $\mathbf{y}$.

Define the asymmetric distance as $\max(\Delta(\mathbf{u},\mathbf{v}),\Delta(\mathbf{v},\mathbf{u}))$, where $\Delta(\mathbf{u},\mathbf{v}) = \sum_i \max(u_i - v_i, 0)$.

Code Design Criteria



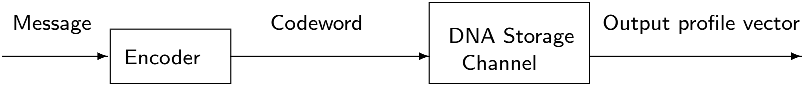
$$
\begin{array}{rcrrrrrrrr}
 & & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\
\mathbf{p}(\mathbf{x}; 2, 3) & = & (0, & 2, & 0, & 1, & 2, & 0, & 1, & 0) \\
\mathbf{p}(\mathbf{y}; 2, 3) & = & (0, & 0, & 3, & 0, & 0, & 3, & 0, & 0)
\end{array}
$$

Criterion 2

Codewords whose ℓ-substrings are resilient to errors.

Certain reliability considerations in DNA storage sequence designs:

- Balanced profiles of ℓ-substrings. Number of C, G bases needs to be roughly fifty percent.
- Forbidden ℓ-substrings. Certain substrings like GCG and CGC or GGG are more likely to cause sequencing errors.



$$\left\{ \begin{array}{c} \text{Yes} \\ \text{No} \end{array} \right\} \qquad \left\{ \begin{array}{l} \mathbf{x} = 10011001 \\ \mathbf{y} = 10101010 \end{array} \right\} \qquad \hat{\mathbf{x}} = (0,1,0,2,0,1,1,0)$$

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

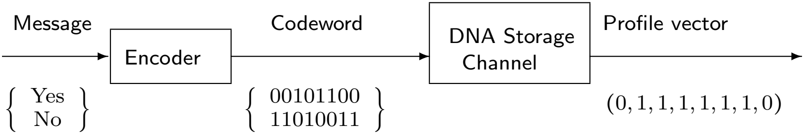
Criterion 2

Codewords whose ℓ-substrings are resilient to errors.

Here, we assume that the ℓ-substrings belong to
$S = \{001, 010, 011, 100, 101, 110\}$.

Fundamental Question - Number of Distinct Profile Vectors

The following example is bad, because the codewords share the same profile vector.



Distinct ℓ-gram Profile Vectors

Define $\mathcal{Q}(n;S)$ to be the set of q-ary words of length n with distinct ℓ-gram profile vectors whose ℓ-grams belong to S.
Determine the size of $\mathcal{Q}(n;S)$.

- n : length of codewords
- q : alphabet size
- ℓ : length of substrings / grams
- S : set of "constrained" substrings (note S is a set of q-ary strings of length ℓ)

Fundamental Question - Code Construction

ℓ-gram Reconstruction Code (GRC)

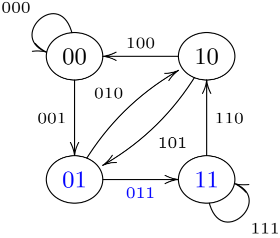
$\mathcal{C} \subseteq \mathcal{Q}(n;S)$ is an $(n,d;S)$-ℓ-GRC if the ℓ-gram distance between any pair of distinct words is at least d.
Construct good $(n,d;S)$-ℓ-GRC. Good means "more codewords."

- n : length of codewords
- q : alphabet size
- ℓ : length of substrings / grams
- S : set of "constraint" substrings (note S is a set of q-ary strings of length ℓ)
- d : minimum ℓ-gram distance between any pair of codewords

Enumeration of Profile Vectors

De Bruijn Graphs

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



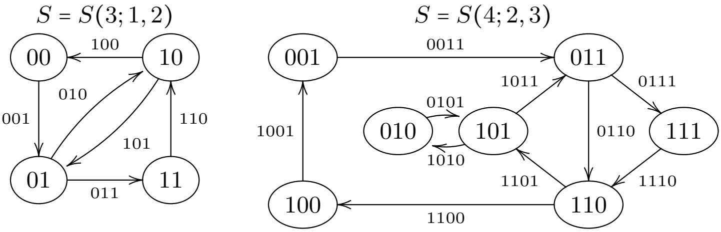
De Bruijn Graphs (de Bruijn, 1946)

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

$$\begin{array}{ccccc} v_2 & v_3 & & v_{\ell-1} & \\ \shortparallel & \shortparallel & \cdots & \shortparallel & . \\ v'_1 & v'_2 & & v'_{\ell-2} & \end{array}$$

Restricted De Bruijn Graphs

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

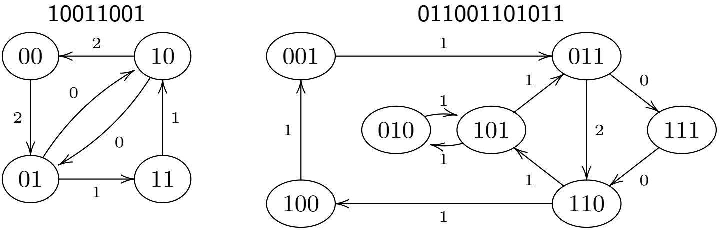


Restricted de Bruijn Graphs $D(S)$ (Ruskey, Sawada, Williams, 2012)

Nodes V are $\ell - 1$-prefixes and -suffixes of strings in S.
$(\mathbf{v}, \mathbf{v}')$ is an arc if

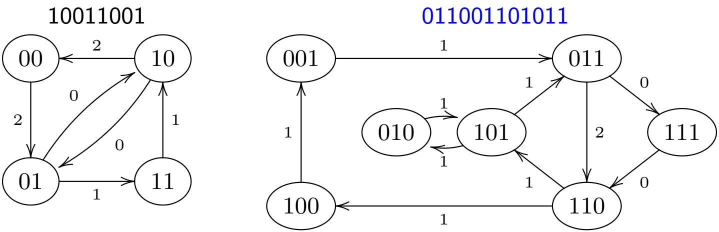
$$\begin{array}{cccc} v_2 & v_3 & & v_{\ell-1} \\ \shortparallel & \shortparallel & \dots & \shortparallel \\ v'_1 & v'_2 & & v'_{\ell-2} \end{array} \quad \text{and} \quad v_1 v_2 \cdots v_{\ell-1} v'_{\ell-1} \in S.$$

Profile Vectors and Flow Vectors



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

Profile Vectors and Flow Vectors



Profile vectors of closed words in $\mathcal{Q}(n;S)$ are flow vectors in $D(S)$.

Closed Words

Closed words that are words that start and end with the same $(\ell - 1)$-gram. Denote the set of q-ary words of length n with distinct ℓ-gram profile whose ℓ-grams belong to S by $\overline{\mathcal{Q}}(n;S)$.

Flow Vectors

Incoming flow is equal to outgoing flow at each node.

Necessary Conditions

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

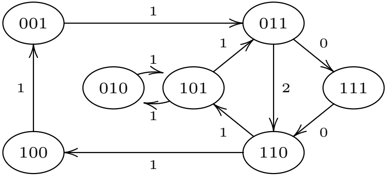
Flow conservations equations:

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

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 $\mathbf{b} = (1, 0, \ldots, 0)^T$. We rewrite the equations as

Necessity

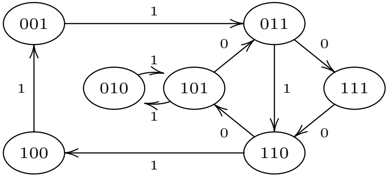
$$\mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b} \text{ and } \mathbf{u} \geq \mathbf{0}.$$

Flow vectors are not always profile vectors

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

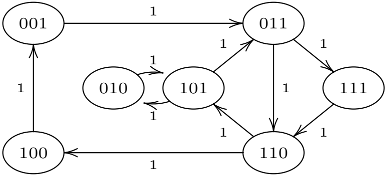
$$\mathbf{A}\mathbf{u} = (n - \ell + 1)\mathbf{b}.$$

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.

Sufficiency

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

Profile Vectors and Lattice Points

Consider the following two sets of lattice points:

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

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

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

Profile Vectors and Lattice Points

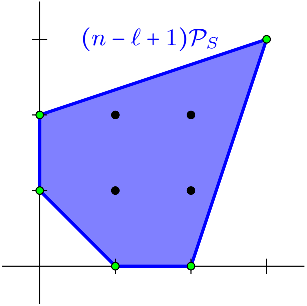
Consider the following two sets of lattice points:

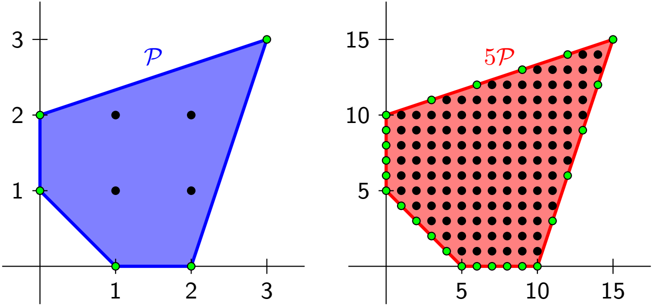
$$\mathcal{F}(n;S) \triangleq \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}\mathbf{u} = (n-\ell+1)\mathbf{b},\ \mathbf{u} \geq \mathbf{0}\},$$
$$\mathcal{E}(n;S) \triangleq \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}\mathbf{u} = (n-\ell+1)\mathbf{b},\ \mathbf{u} > \mathbf{0}\}.$$

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

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 $(n-\ell+1)\mathcal{P}_S$.





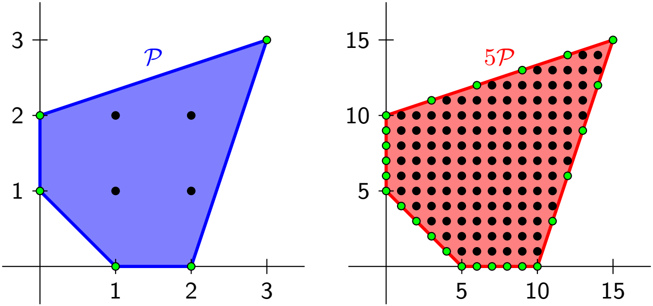
For a polytope $\mathcal{P} \subset \mathbb{R}^N$ and $t \in \mathbb{R}$, the dilation $t\mathcal{P}$ is given by

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

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

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

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

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

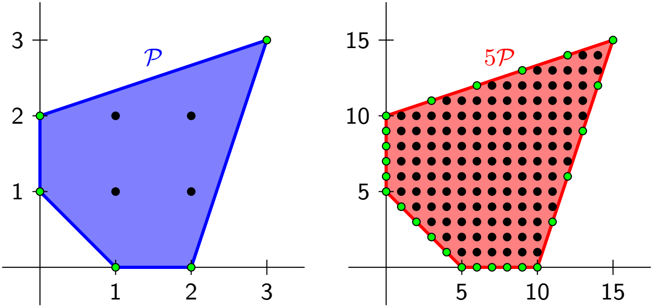
The lattice point enumerator for $\mathcal{P}$ is $\mathcal{L}_\mathcal{P} : \mathbb{R} \to \mathbb{Z}$ defined by

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

Theorem (Ehrhart)

If $\mathcal{P}$ is a rational D-dimensional polytope, then $\mathcal{L}_\mathcal{P}(t)$ is a "quasipolynomial" in t with 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

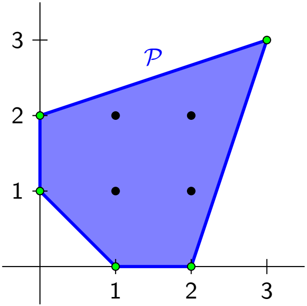
$$t\mathcal{P} = \{tx : x \in \mathcal{P}\}.$$

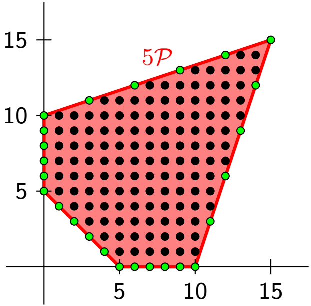
The lattice point enumerator for $\mathcal{P}$ is $\mathcal{L}_{\mathcal{P}} : \mathbb{R} \to \mathbb{Z}$ defined by

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

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" with degree D.





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

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

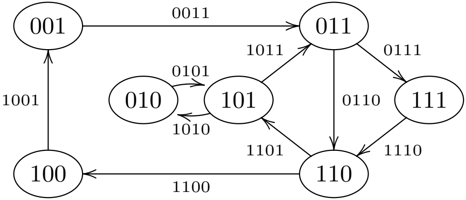
The lattice point enumerator for $\mathcal{P}$ is $\mathcal{L}_{\mathcal{P}} : \mathbb{R} \to \mathbb{Z}$ defined by

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

Lemma

The polytope $\mathcal{P}_S$ has dimension $|S| - |V|$ if $D(S)$ is strongly connected.

Main Enumeration Result



Here, $|S| = 10$, $|V| = 7$ and so, the number of distinct profile vectors of closed words is

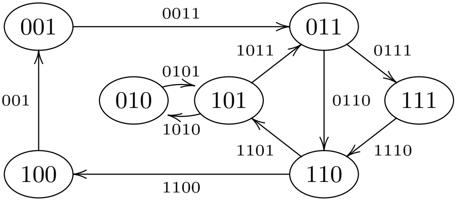
$$|\overline{\mathcal{Q}}(n;S)| = \Theta'(n^3).$$

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

- A *quasipolynomial* f is a function in n of the form $c_D(n)n^D + c_{D-1}(n)n^{D-1} + \cdots + c_0(n)$, where $c_D, c_{D-1}, \ldots, c_0$ are periodic functions in n. If c_D is not identically equal to zero, f is said to be of *degree* D.
- $f(n) = \Omega'(g(n))$ 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 $\lambda|(n - \ell + 1)$. Furthermore, $f(n) = \Theta'(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega'(g(n))$.

Main Enumeration Result



Here, $|S| = 10$, $|V| = 7$ and so, the number of distinct profile vectors of closed words is

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

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

Results hold if n satisfies certain periodicity conditions.

Code Constructions

Varshamov Codes

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

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

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

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

Theorem (Varshamov, 1973)

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

GRC as Subcodes of Varshamov Codes

Construction I

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

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

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

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

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

of which, three are profile vectors in $\mathbf{p}\mathcal{Q}(8;S)$ (profile vectors of words of length eight).

Construction I

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

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

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

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

$(4,0,0,1,0,1)$			$(0,1,1,4,0,0)$		
$(2,2,0,2,0,0)$	$\leftrightarrow$	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
$(1,0,1,0,0,4)$					

of which, three are profile vectors in $\mathbf{p}\mathcal{Q}(8;S)$ (profile vectors of words of length eight).

How many codewords does Construction I guarantee?

Ehrhart Theory Continues

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

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

Proposition

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

$$\mathcal{P}_{\mathrm{GRC}} = \left\{\mathbf{u} \in \mathbb{R}^{|S|+d} : \mathbf{A}_{\mathrm{GRC}}\mathbf{u} = (n-\ell+1)\mathbf{b},\ \mathbf{u} \geq \mathbf{0}\right\}.$$

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

Theorem

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

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

- $f(n) = \Omega'(g(n))$ 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 $\lambda|(n-\ell+1)$.

Systematic Encoding

Objective

Efficient one-to-one mapping

$$\phi : \{0, 1, \ldots, m-1\}^{|S|-|V|-1} \to \mathbf{p}\mathcal{Q}(n; S)$$

such that $\mathbf{v}$ is "embedded" in $\phi(\mathbf{v})$.

For example, 012 encodes systematically into $(3, 1, 0, 2, 1, 1, 2, 2)$, the profile vector of 00000110111100.

Systematic Encoding

Objective

Efficient one-to-one mapping

$$\phi : \{0, 1, \ldots, m-1\}^{|S|-|V|-1} \to \mathbf{p}\mathcal{Q}(n; S)$$

such that $\mathbf{v}$ is "embedded" in $\phi(\mathbf{v})$.

For example, 012 encodes systematically into $(3, 1, 0, 2, 1, 1, 2, 2)$, the profile vector of 00000110111100.

Theorem

Suppose $D(S)$ is Hamiltonian and contains loops. For a suitable choice of m, we can systematically encode $\{0, 1, \ldots, m-1\}^{|S|-|V|-1}$ into $\mathbf{p}\mathcal{Q}(n; S)$.

Systematic Encoding

Construction II

Suppose $D(S)$ is Hamiltonian and contains loops. For a suitable choice of m, if $\mathcal{C}$ is an m-ary $(|S|-|V|-1,d)$-AECC, then $\{\phi(\mathbf{v}) : \mathbf{v} \in \mathcal{C}\}$ is a $(n,d;S)$-ℓ-GRC.

Consider the following code with minimum asymmetric distance 3.

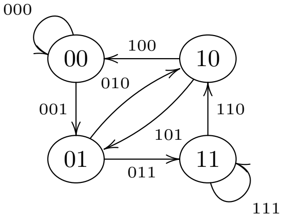
$$\{(0,0,0),(1,4,2),(2,3,4),(3,2,1),(4,1,3)\}.$$

We systematic encode them to 3-gram profile vectors of words of length 8.

$(0,0,0)$	$(0,0,1)$	$(0,1,0)$	$(0,1,1)$	$(1,0,0)$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$
$(18,$	$0,$	$0,$	$0,$	$0,$	$0,$	$0,$	$0)$
$(1,$	$1,$	$1,$	$4,$	$1,$	$4,$	$4,$	$2)$
$(3,$	$1,$	$2,$	$2,$	$1,$	$3,$	$2,$	$4)$
$(6,$	$2,$	$3,$	$1,$	$2,$	$2,$	$1,$	$1)$
$(0,$	$4,$	$4,$	$1,$	$4,$	$1,$	$1,$	$3)$

This forms a 3-gram reconstruction code of length 20 and distance at least 3.

Corollaries of Main Enumeration Result



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

Theorem (Jacquet, Knessl, Szpankowski, 2012)

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

Corollary

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

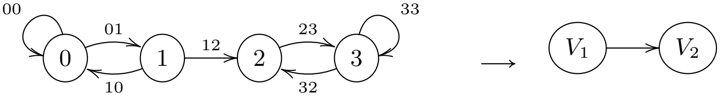
$$|\mathcal{E}(n;S| \sim |\mathcal{F}(n;S)| \sim |\overline{\mathcal{Q}}(n;S)| \sim c(S)n^{|S|-|V|} \text{ where } c(S) \text{ is a constant.}$$

Corollaries of Main Enumeration Result

Results can be extended to enumerate profile vectors of

- all words (not nec. closed) with $D(S)$ strongly connected;
- closed words with $D(S)$ not strongly connected;
- all words with $D(S)$ not strongly connected.

$q = 4,\ \ell = 2,\ S = \{00, 01, 10, 12, 23, 32, 33\}$



Proof Idea when $D(S)$ is not strongly connected: Consider the strongly connected components of $D(S)$ and apply the main enumeration result.

Computing the Leading Coefficient for $\mathcal{Q}(n;S)$

Recall the polytope $\mathcal{P} = \{\mathbf{u} \in \mathbb{R}^{|S|} : \mathbf{A}\mathbf{u} = (n-\ell+1)\mathbf{b},\ \mathbf{u} \geq \mathbf{0}\}$.
The lattice point enumerator

$$L(n-\ell+1) = \#(\mathbb{Z}^n \cap (n-\ell+1)\mathcal{P})$$

gives the number of points in $\mathcal{F}(n;S)$. Hence, the "leading coefficient" for $\mathcal{Q}(n;S)$.
The lattice point enumerator can be computed in polynomial time when the dimension of the polytope is fixed (Barvinok, 1994). However, the dimension of $\mathcal{P}$ is $|S| - |V| \approx q^{\ell}$.

Question

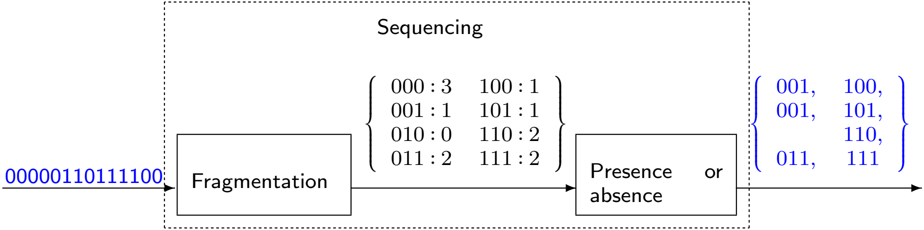
Efficient methods to compute the lattice point enumerator or compute the leading coefficient.

Practical Methods for Profile Counting

Challenge: Counting accurately the number of ℓ-grams.
Instead, we use certain auxiliary information.

- The presence or absence of ℓ-grams
- The relative order of ℓ-grams

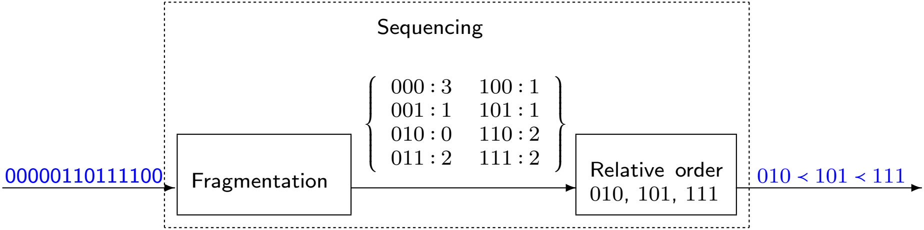
Practical Methods for Profile Counting - Presence or Absence of ℓ-grams



Related combinatorial problems:

- Enumeration of "profile vectors". Tan, Shallit (2013) studied this problem in the context of "factors of words".
- Edge-disjoint path decompositions of de Bruijn graphs. Variety of decompositions surveyed by Heinrich (1993), Bryant and El-Zanati (2007). Cooper and Graham (2004) studied cycle decompositions of de Bruijn graphs.

Practical Methods for Profile Counting - Relative Order of ℓ-grams



Related coding problem:

- Rank modulation codes. Jiang, Mateescu, Schwartz, Bruck (2009) proposed these codes for nonvolatile flash memories.

Growth rate for $\mathcal{Q}(n;S)$ when ℓ grows

When q, ℓ is fixed, $|\mathcal{Q}(n;S)|$ is polynomial in n.
Suppose that q is fixed and ℓ is a function of n, or $\ell = f(n)$.

- For example, when $\ell = n$ and $S = \{0, 1, \ldots, q-1\}^{\ell}$, then $|\mathcal{Q}(n;S)| = q^n$, which has exponential growth in n.

Question

How "small" can ℓ be so as to ensure $|\mathcal{Q}(n;S)|$ has exponential growth in n?

Encoding and Decoding Complexities

Recall that n is the length of codewords.

ℓ-gram distance and code constructions are defined using profile vectors of length $|S| \approx q^{\ell}$.

When $n \le q^{\ell}$, computations based on profile vectors are inefficient.

Ukkonen (1992) showed that (a variant) of the ℓ-gram distance can computed in time $O(qn)$ with space $O(qn)$.

Question

Assume q fixed. Can encoding and decoding by done in time and space polynomial in n?

Acknowledgment

Collaborators:
Han Mao Kiah
Gregory Puleo

Funding:
CIA
UIUC SRI
NSF
DARPA Molecular Informatics