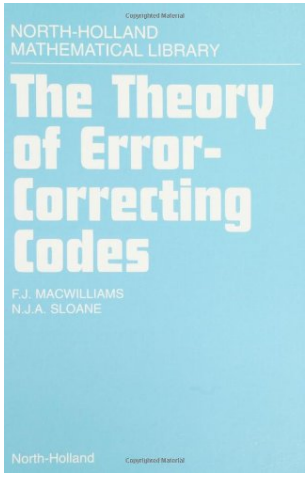
Codes for DNA sequence profiles

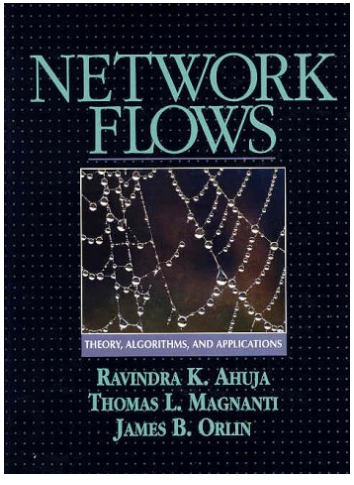
H. M. Kiah

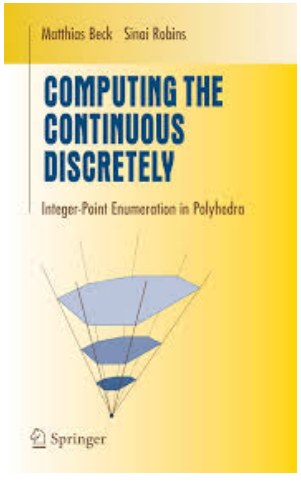
Joint work with G. Puleo (UIUC) and O. Milenkovic (UIUC)

Nanyang Technological University, Singapore

ISIT, 16 Jun 2015







"Double Helix Serves Double Duty", NY Times, Jan 2013

Digital information storage in synthetic DNA:

- Goldman *et al.* (Nature, 2013) stored 739 KB of data on synthetic DNA, shipped it from USA to Germany and recreated the original digital files "without errors".
- "a step towards digital archival storage medium of immense scale".
- Goal: to store the equivalent of one million CDs in a gram of DNA for 10,000 years.

Neanderthal extinction: 35,000 years ago - DNA is extremely durable!

DNA Synthesis and Sequencing

Central to a DNA information storage: DNA synthesis and sequencing.

- DNA synthesis refers to the "write" process.
- DNA sequencing refers to the "read" process.
- Both involve complex biochemical processes, with costs decreasing daily.

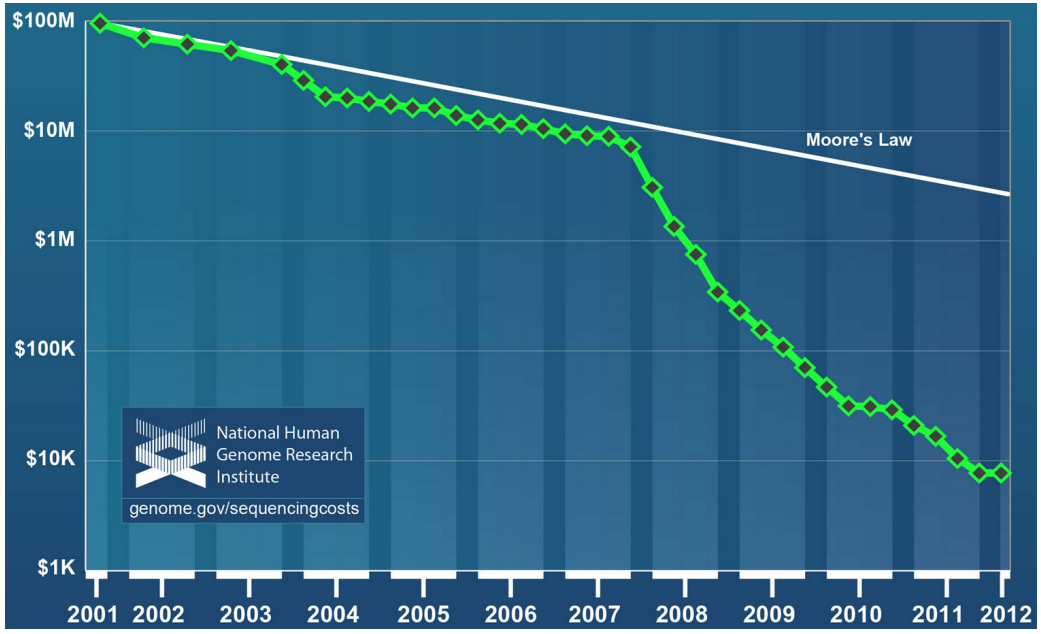
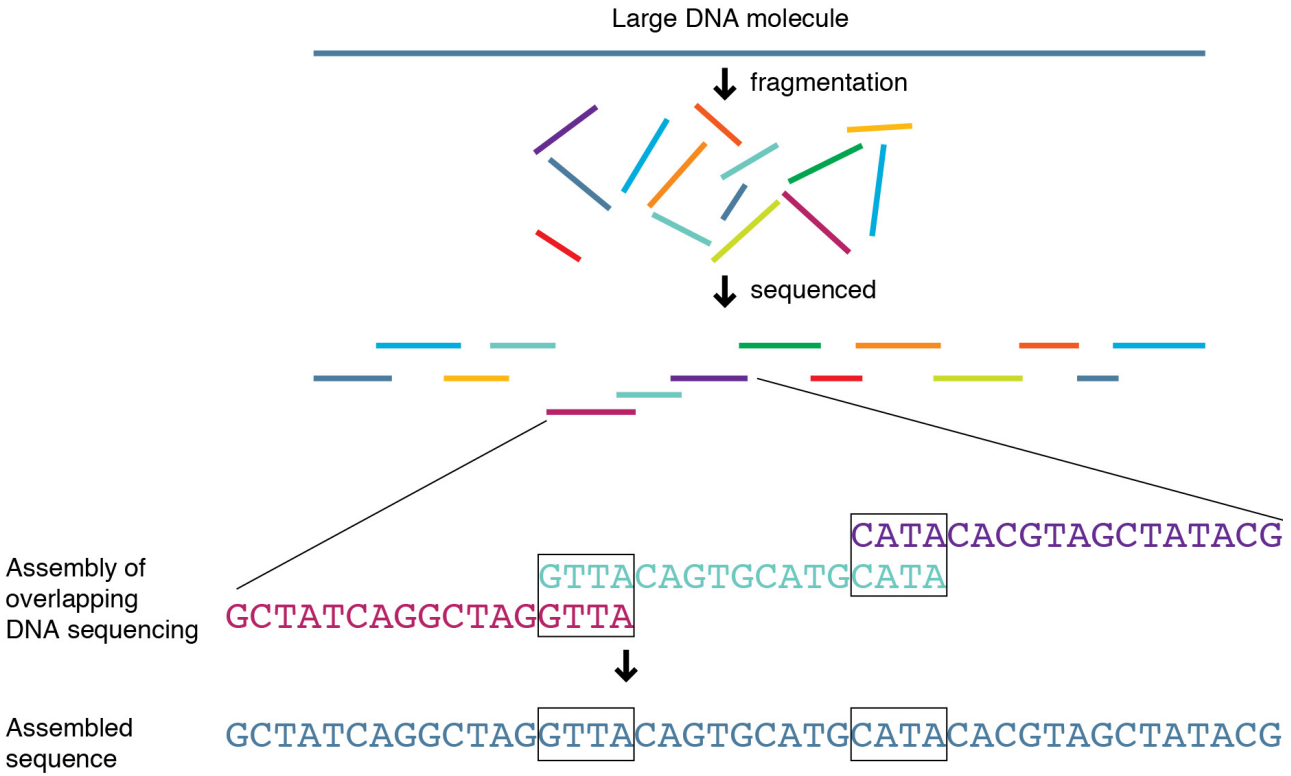


Figure: Cost of sequencing a genome

Sequence Assembly Problem

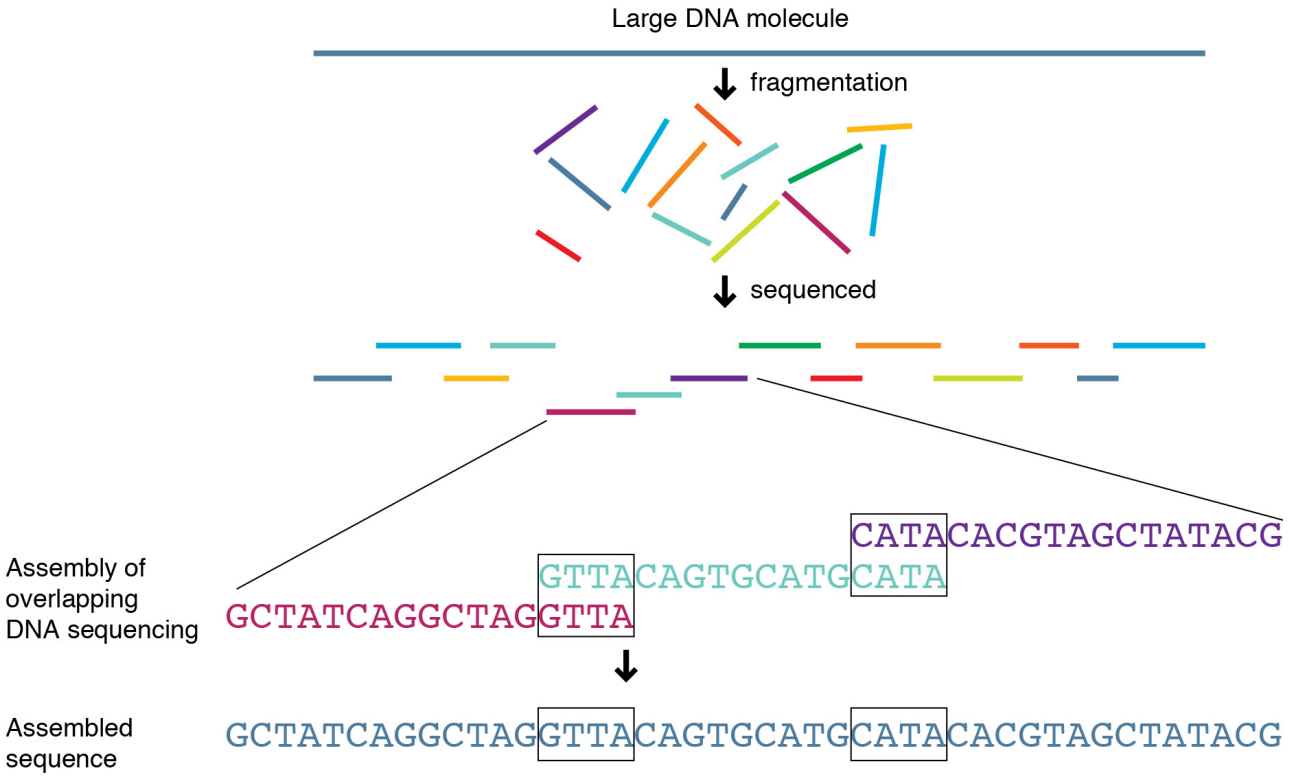
Sequencing is computationally demanding.



Need to stitch together many short reads to obtain original sequence.

Sequence Assembly Problem

Sequencing is computationally demanding.

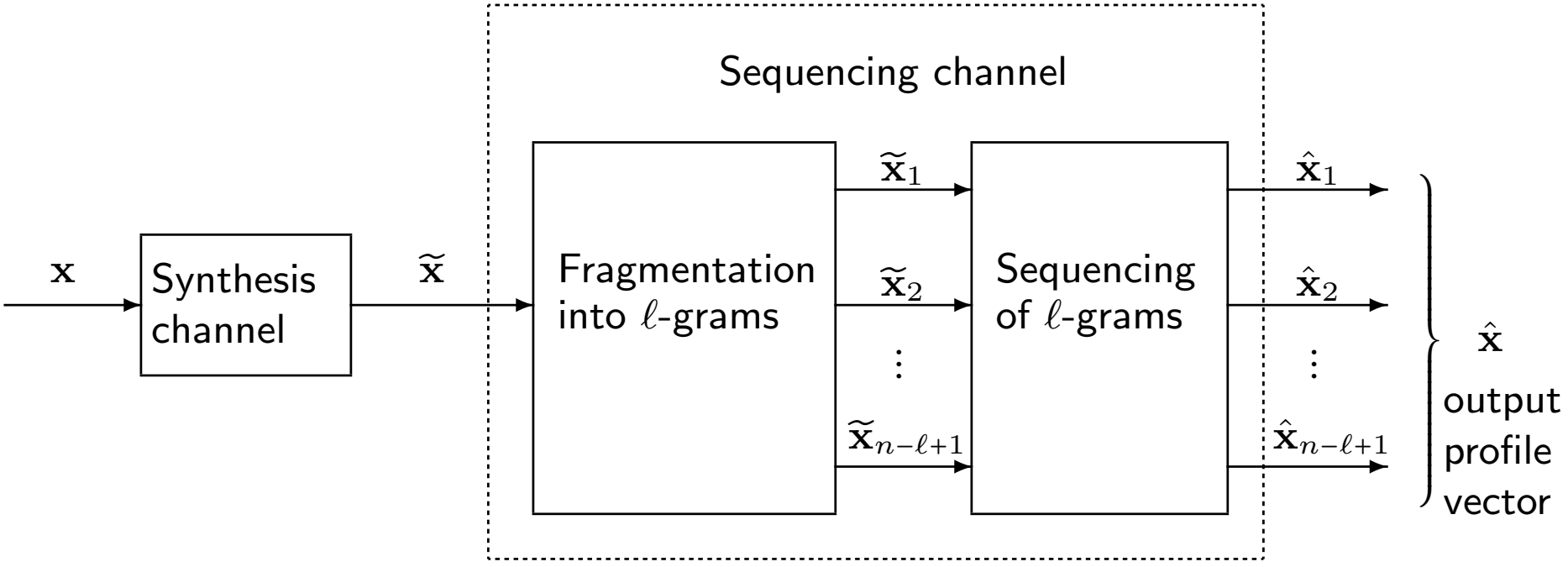


Need to stitch together many short reads to obtain original sequence.

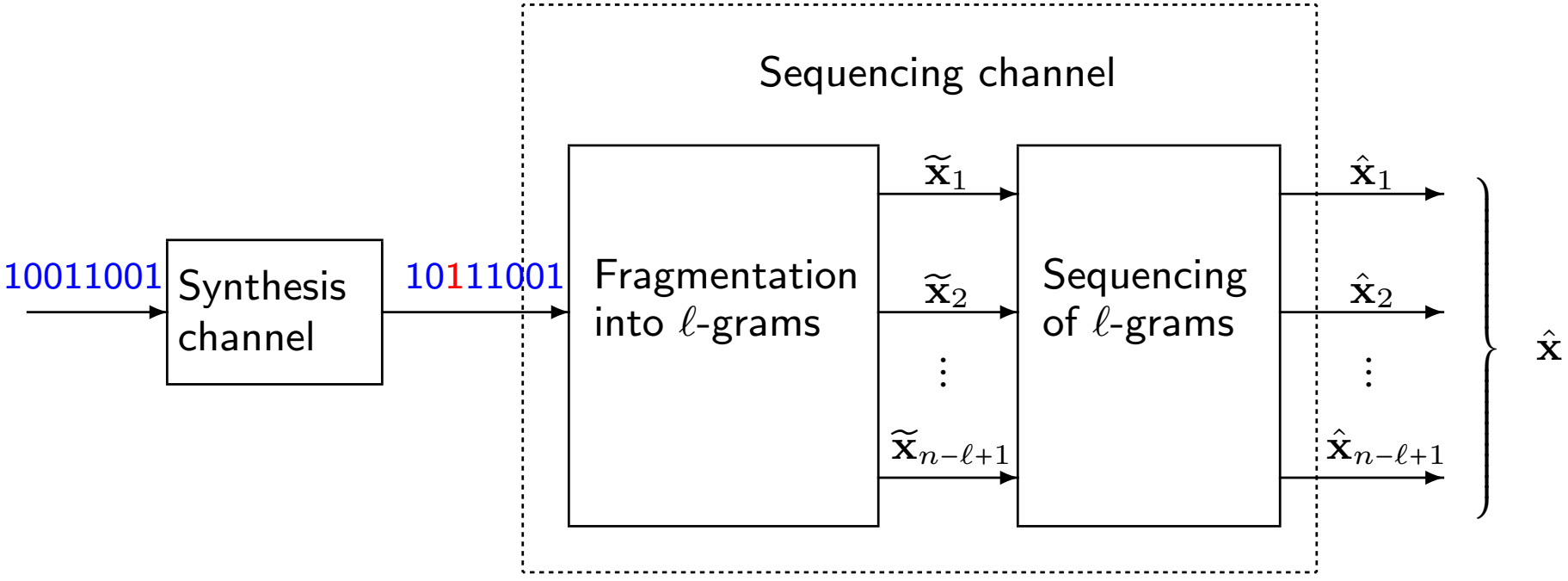
Idea

Design a code that uses the information on short reads / substrings directly, without the need to stitch them together.

DNA Storage Channel: A (Slight) Abstraction



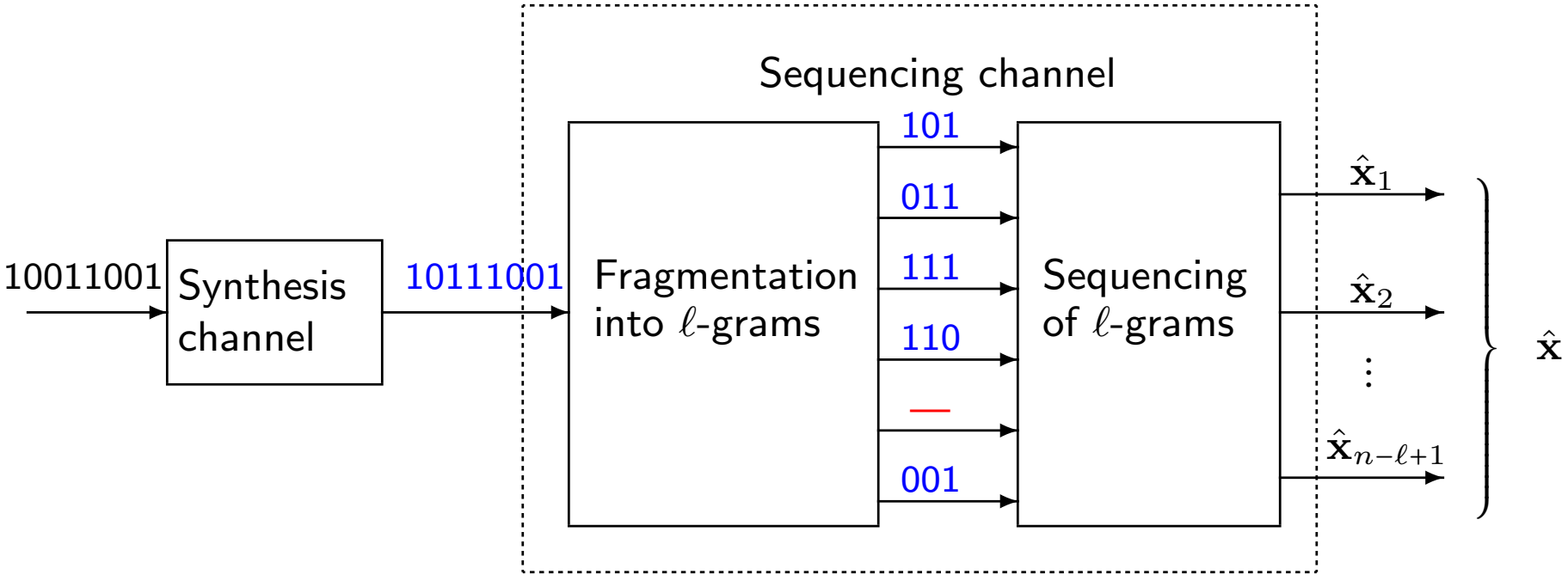
DNA Storage Channel



Synthesis channel captures the "write" process.

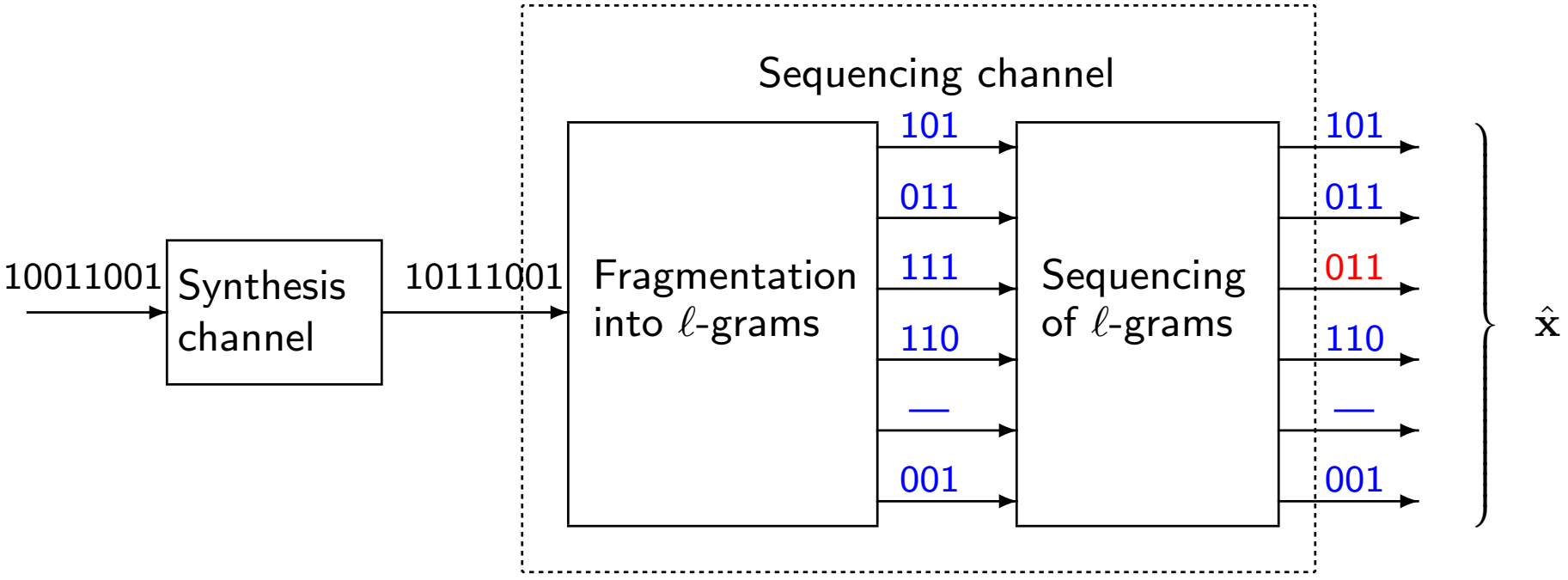
The sequence synthesis process introduces errors (current technologies $\leq 1\%$).

DNA Storage Channel



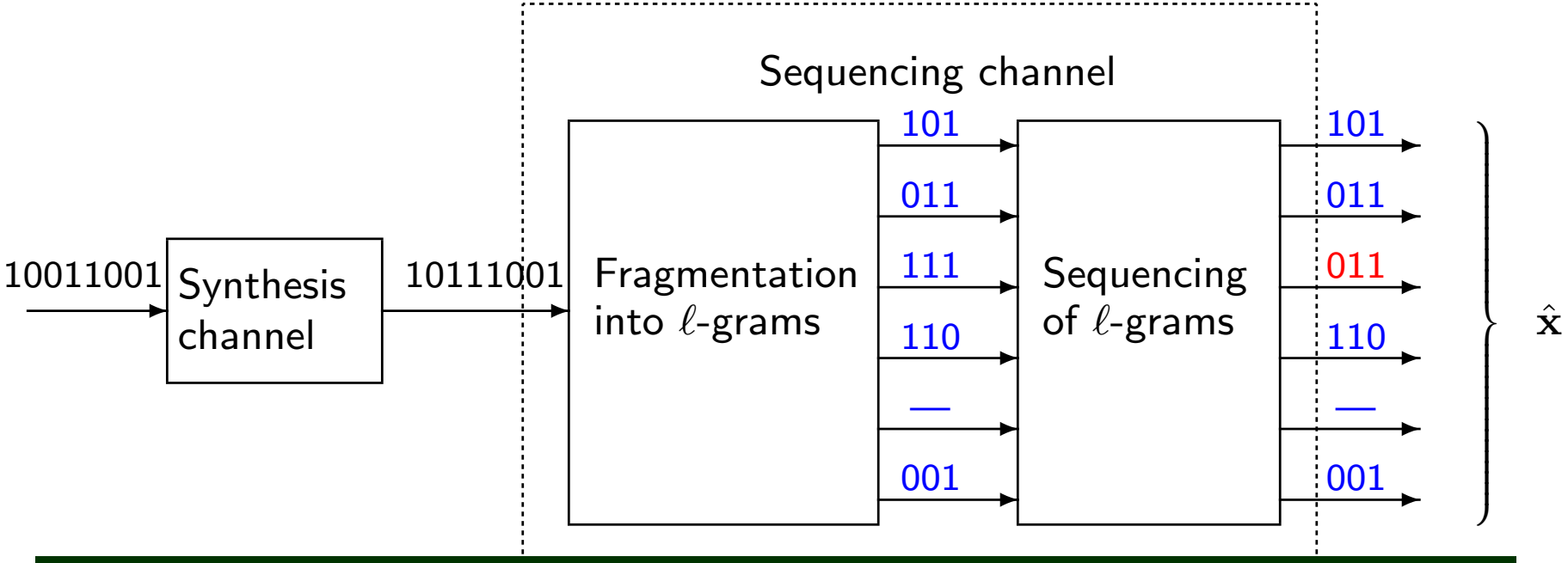
DNA sequencing represents the "read" process.
DNA sequencing is technologically more advanced and cheaper than synthesis, but coupled with computational difficulties.

DNA Storage Channel



Sequencing also introduces errors in fragments (reads) (current Illumina platforms have error rate $\leq 1\%$).

DNA Storage Channel: Profiles



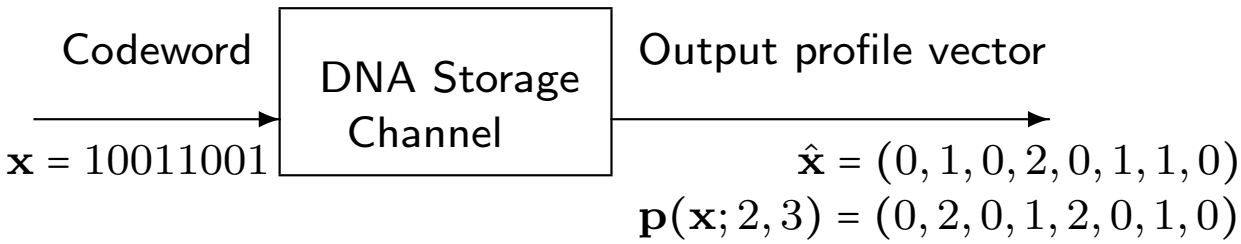
Output profile vector

Given an input sequence 10011001, we obtain an output profile vector that reflects the count of each substring at the channel output:

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

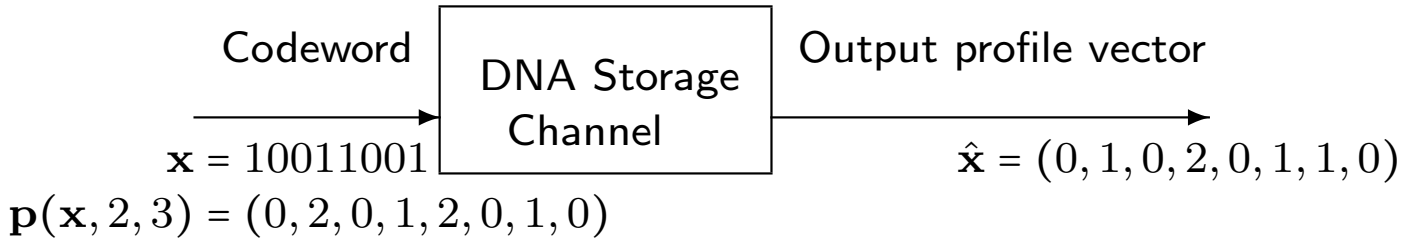
Note: position of substring is not known!

Profile Vectors



Profile vector is what the storage channel outputs when there is no error.

Profile Vectors



Profile Vector

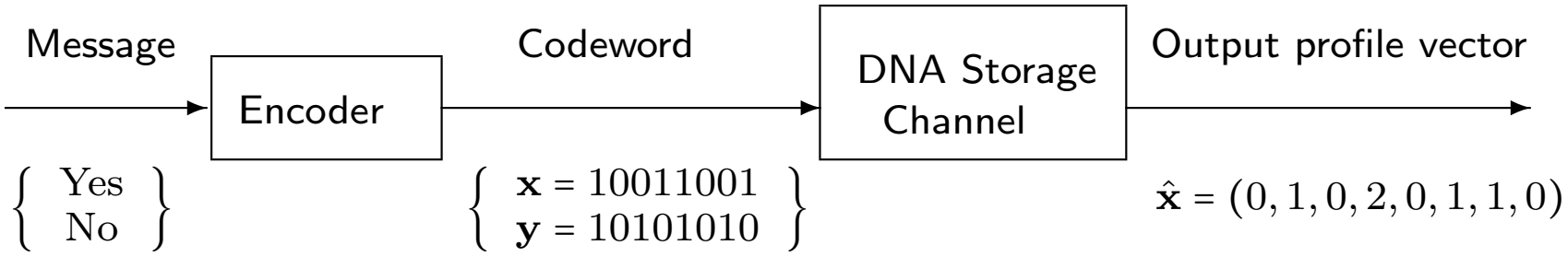
Fix q and $\ell < n$. Let $\mathbf{p}(\mathbf{x}; q, \ell)$ denote the profile vector indexed by $[q]^\ell$, where the entry for the ℓ-gram $\mathbf{z}$ gives the number of occurrences of $\mathbf{z}$ in $\mathbf{x}$.

Example

Given $\mathbf{x} = 10011001$, then $\mathbf{p}(\mathbf{x}; 2, 3) =$

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

Code Design Criteria



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

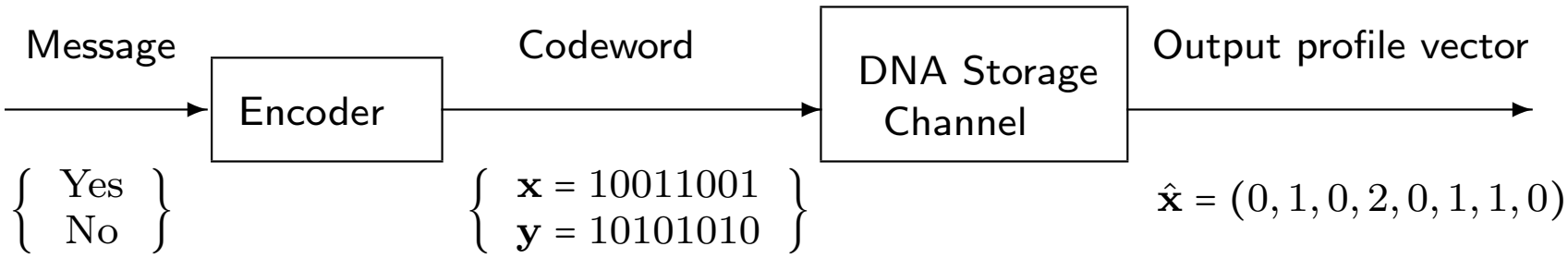
Criterion 1: Error-control

Codewords whose profile vectors are "far from each other".

We define the ℓ-gram distance between $\mathbf{x}$ and $\mathbf{y}$ as the asymmetric distance between $\mathbf{p}(\mathbf{x};2,3)$ and $\mathbf{p}(\mathbf{y};2,3)$.

Asymmetric distance: $\max(\Delta(\mathbf{u},\mathbf{v}),\Delta(\mathbf{v},\mathbf{u}))$, where $\Delta(\mathbf{u},\mathbf{v}) = \sum_i \max(u_i - v_i, 0)$.

Code Design Criteria



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

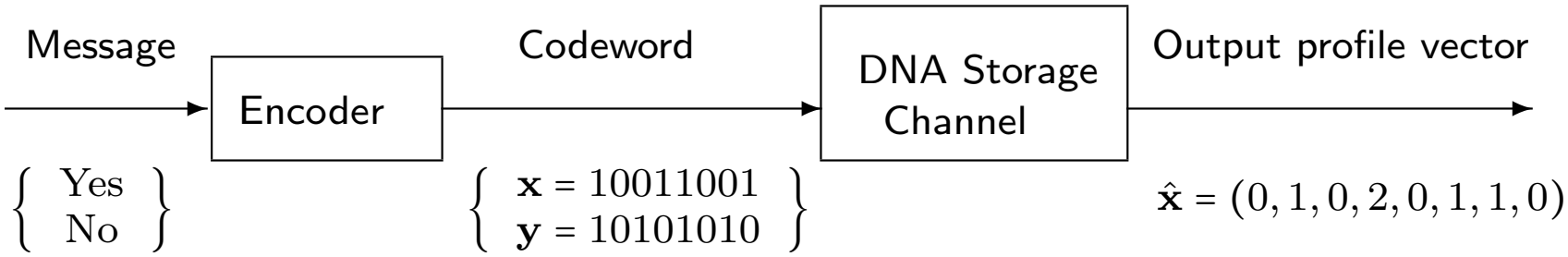
Criterion 2: Constrained Coding

Codewords whose ℓ-substrings are resilient to errors.

Certain reliability considerations in DNA storage sequence designs:

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

Code Design Criteria



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

Criterion 2: Constrained Coding

Codewords whose ℓ-substrings are resilient to errors.

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

Fundamental Questions

Distinct ℓ-gram Profile Vectors

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

Note: 00101100 and 11010011 have the same profile vector for $\ell = 3$.

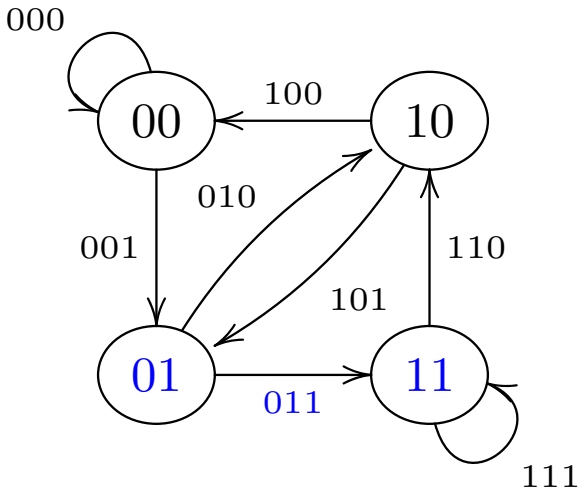
ℓ-gram Reconstruction Code (GRC)

$\mathcal{C} \subseteq \mathcal{Q}(n; S)$ 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.

- 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 of a code

De Bruijn Graphs

Here, $q = 2$, $\ell = 3$.



De Bruijn Graphs (de Bruijn, 1946)

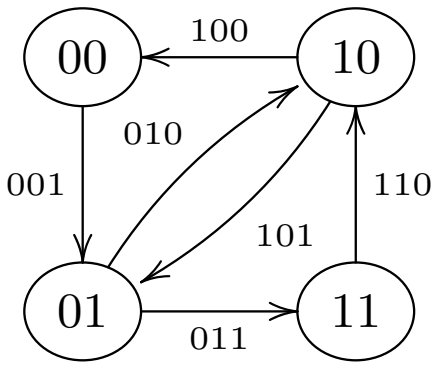
Nodes are q-ary strings of length $\ell - 1$.
$(\mathbf{v}, \mathbf{v}')$ is an arc if

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

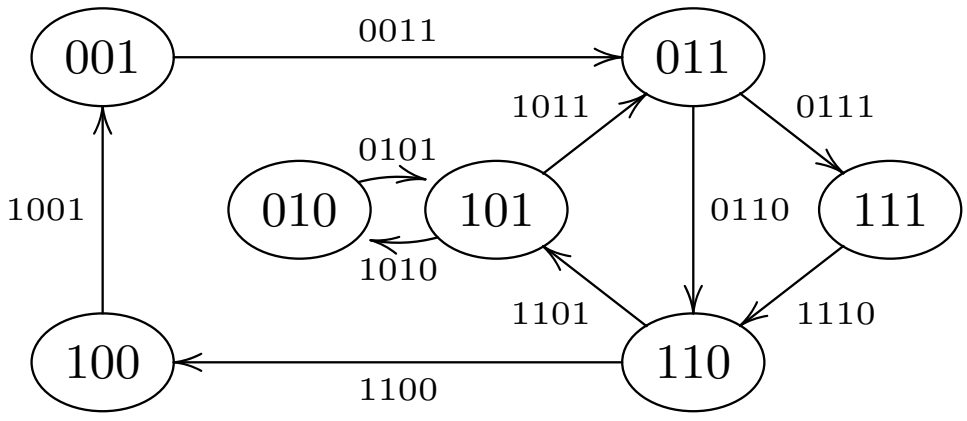
Restricted De Bruijn Graphs

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

$S = S(3; 1, 2)$



$S = S(4; 2, 3)$

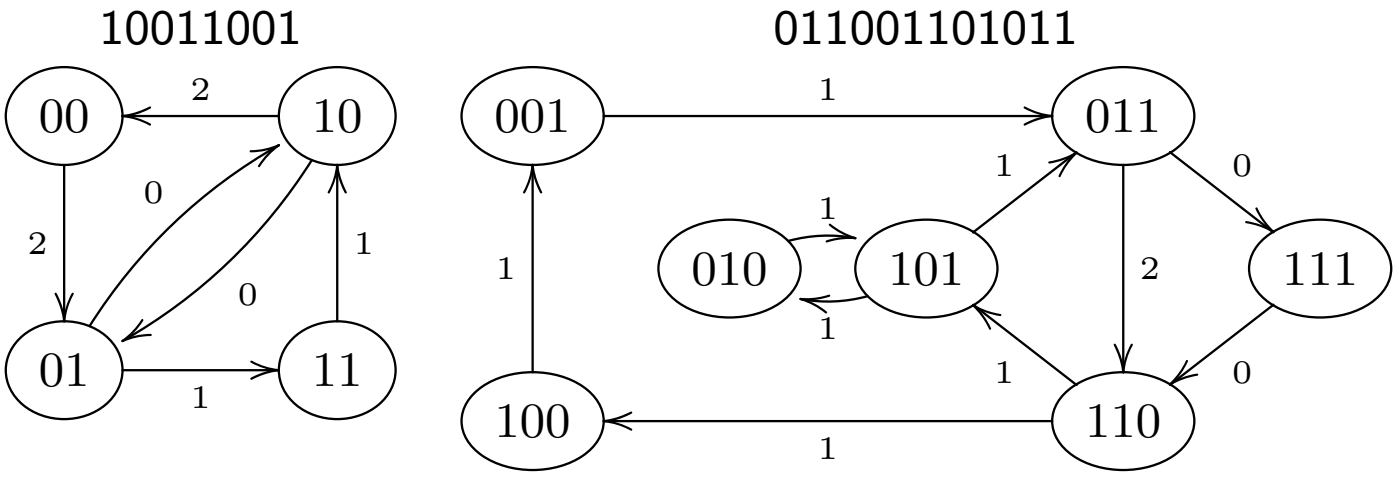


Restricted de Bruijn Graphs $D(S)$ (Ruskey, Sawada, Williams, 2012)

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

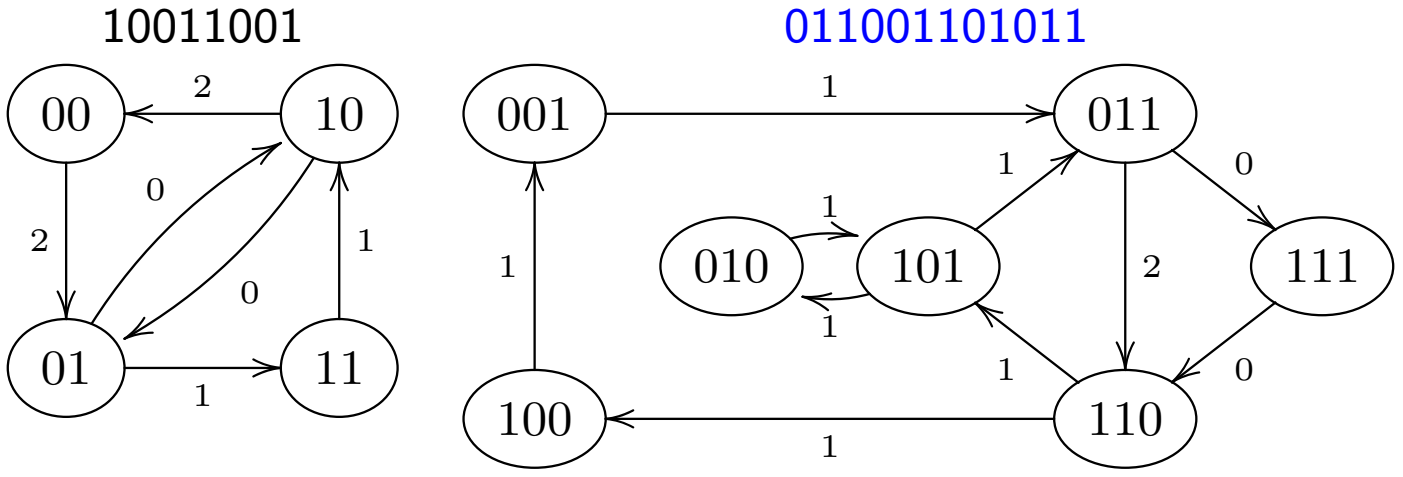
$$\begin{array}{cccccc} v_2 & v_3 & & v_{\ell-1} & & \\ \| & \| & \ldots & \| & \text{and} & v_1 v_2 \cdots v_{\ell-1} v'_{\ell-1} \in S. \\ v'_1 & v'_2 & & v'_{\ell-2} & & \end{array}$$

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



A closed word is a word that begins and ends with the same $(\ell-1)$-gram. Profile vectors of closed words in $\mathcal{Q}(n;S)$ are flow vectors in $D(S)$: at each node, total incoming flow = total outgoing flow.

Idea: to count words (up to ℓ-gram equivalence), count integer flow vectors instead.
(Not all flow vectors correspond to closed words, but asymptotically this works.)

Necessary Conditions

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

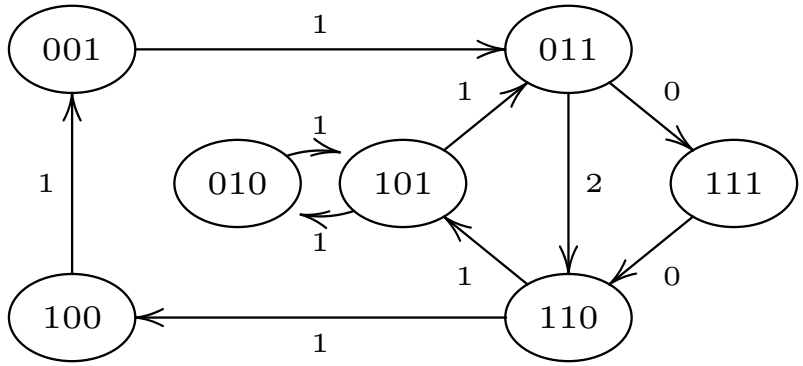
Flow conservation:

$$\mathbf{Bu} = \mathbf{0},$$

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

Sum of flows:

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



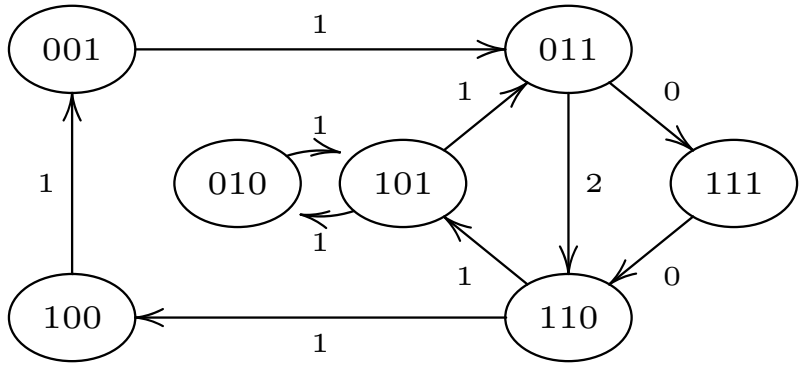
Necessary Conditions

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

Flow conservation:

$$\mathbf{Bu} = \mathbf{0},$$

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



Sum of flows:

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

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

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

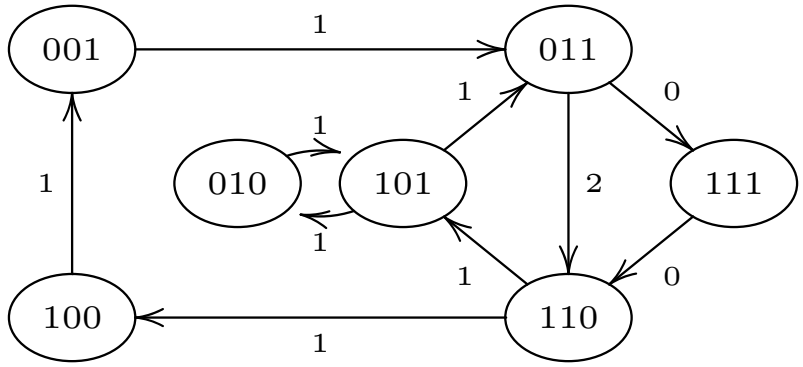
Necessary Conditions

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

Flow conservation:

$$\mathbf{Bu} = \mathbf{0},$$

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



Sum of flows:

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

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

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

Thus, flow vectors for $D(S)$ correspond to integer points in the following polytope:

$$\mathcal{P}_{n-\ell+1} = \{\mathbf{u} \in \mathbb{R}^S : \mathbf{u} \geq 0,\ \mathbf{Au} = (n - \ell + 1)\mathbf{b}\}$$

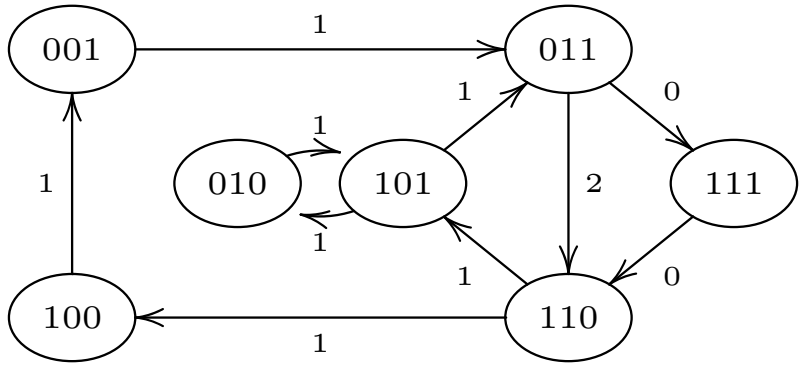
Necessary Conditions

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

Flow conservation:

$$\mathbf{Bu} = \mathbf{0},$$

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



Sum of flows:

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

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

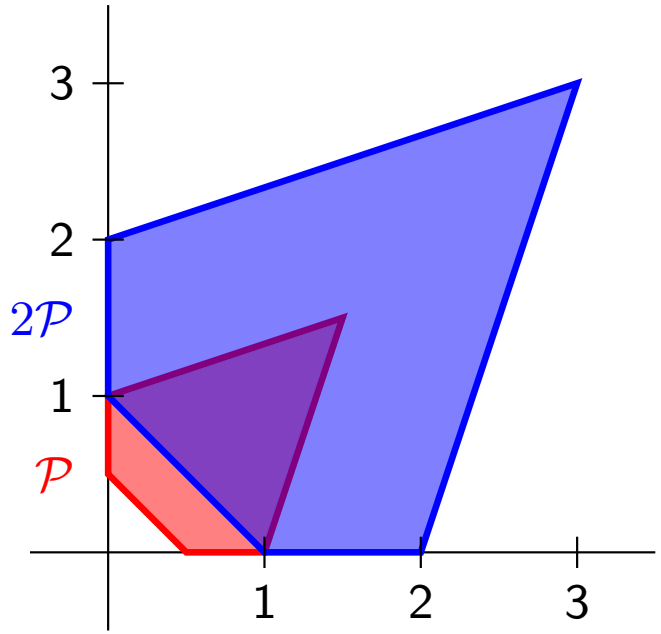
$$\mathbf{Au} = (n - \ell + 1)\mathbf{b} \text{ and } \mathbf{u} \geq \mathbf{0}.$$

Thus, flow vectors for $D(S)$ correspond to integer points in the following polytope:

$$\mathcal{P}_{n-\ell+1} = \{\mathbf{u} \in \mathbb{R}^S : \mathbf{u} \geq 0,\ \mathbf{Au} = (n - \ell + 1)\mathbf{b}\}$$

We rephrase this in terms of dilating a fixed polytope $\mathcal{P}$.

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

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

Theorem (Ehrhart)

If $\mathcal{P}$ is a rational polytope, then $\mathcal{L}_{\mathcal{P}}$ is a "quasipolynomial" in t.
In particular, if $\mathcal{P}$ is k-dimensional, then $\mathcal{L}_{\mathcal{P}}(t) = \Theta(t^k)$.

Dilation and Word Length

Flow vectors for $D(S)$ correspond to integer points in the following polytope:

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

Dilation and Word Length

Flow vectors for $D(S)$ correspond to integer points in the following polytope:

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

In particular, $\mathcal{P}_{n-\ell+1} = (n-\ell+1)\mathcal{P}_1$.
Thus, increasing the word length n corresponds to dilating $\mathcal{P}_1$, a fixed polytope.

Dilation and Word Length

Flow vectors for $D(S)$ correspond to integer points in the following polytope:

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

In particular, $\mathcal{P}_{n-\ell+1} = (n-\ell+1)\mathcal{P}_1$.
Thus, increasing the word length n corresponds to dilating $\mathcal{P}_1$, a fixed polytope.

Lemma

If $D(S)$ is strongly connected, then $\dim(\mathcal{P}_1) = |S| - |V(S)|$.
In particular, if S is all q-ary words of length ℓ, then $\dim(\mathcal{P}_1) = q^\ell - q^{\ell-1}$.

Dilation and Word Length

Flow vectors for $D(S)$ correspond to integer points in the following polytope:

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

In particular, $\mathcal{P}_{n-\ell+1} = (n-\ell+1)\mathcal{P}_1$.
Thus, increasing the word length n corresponds to dilating $\mathcal{P}_1$, a fixed polytope.

Lemma

If $D(S)$ is strongly connected, then $\dim(\mathcal{P}_1) = |S| - |V(S)|$.
In particular, if S is all q-ary words of length ℓ, then $\dim(\mathcal{P}_1) = q^\ell - q^{\ell-1}$.

Corollary

If $D(S)$ is strongly connected, $|\mathcal{Q}(n;S)| = \Theta(n^{|S|-|V(S)|})$.
That is, up to ℓ-gram equivalence, there are $\Theta(n^{|S|-|V(S)|})$ words whose ℓ-grams all belong to S.

In the context of Markov types, Jacquet, Knessl, Szpankowski (2012) derived similar results where $S = [q]^\ell$ using different techniques.

Varshamov Codes

- All we've done so far is count words up to ℓ-gram equivalence. (That is, we've enforced an ℓ-gram distance of 1).
- What if we want to force a higher ℓ-gram distance of code words?

Varshamov Codes

- All we've done so far is count words up to ℓ-gram equivalence. (That is, we've enforced an ℓ-gram distance of 1).
- What if we want to force a higher ℓ-gram distance of code words?

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

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

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

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

Theorem (Varshamov, 1973)

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

Codes

- Using Varshamov codes we obtain new $\mathbf{A}$,$\mathbf{b}$ such that when

$$\mathcal{P} = \{\mathbf{u} \in \mathbb{R}^{|S|+k} \colon \mathbf{A}\mathbf{u} = \mathbf{b},\ \mathbf{u} \geq 0\},$$

the integer points of $(n-\ell+1)\mathcal{P}$ correspond to flows in $D(S)$ with sum $n-\ell+1$ whose "profile vectors" are distance $\geq d$ from each other.

Codes

- Using Varshamov codes we obtain new $\mathbf{A}$,$\mathbf{b}$ such that when

$$\mathcal{P} = \{\mathbf{u} \in \mathbb{R}^{|S|+k} \colon \mathbf{A}\mathbf{u} = \mathbf{b},\ \mathbf{u} \geq 0\},$$

  the integer points of $(n-\ell+1)\mathcal{P}$ correspond to flows in $D(S)$ with sum $n-\ell+1$ whose "profile vectors" are distance $\geq d$ from each other.
- If $D(S)$ is strongly connected, still get the same dimension $|S|-|V(S)|$ for this polytope, yielding $\Theta(n^{|S|-|V(S)|})$.

Codes

- Using Varshamov codes we obtain new $\mathbf{A}$,$\mathbf{b}$ such that when

$$\mathcal{P} = \{\mathbf{u} \in \mathbb{R}^{|S|+k} \colon \mathbf{A}\mathbf{u} = \mathbf{b},\ \mathbf{u} \geq 0\},$$

  the integer points of $(n-\ell+1)\mathcal{P}$ correspond to flows in $D(S)$ with sum $n-\ell+1$ whose "profile vectors" are distance $\geq d$ from each other.
- If $D(S)$ is strongly connected, still get the same dimension $|S|-|V(S)|$ for this polytope, yielding $\Theta(n^{|S|-|V(S)|})$.
- Thus, fixing a minimum distance d affects the leading coefficient of the number of code words, but not the exponent.

Questions?

- Details and other results on arXiv.
  - Codes for DNA Sequence Profiles
  - `http://arxiv.org/abs/1502.00517`
- Authors
  - Han Mao Kiah, NTU Singapore, `hmkiah@ntu.edu.sg`
  - Greg Puleo, UIUC USA, `puleo@illinois.edu`
  - Olgica Milenkovic, UIUC USA, `milenkovic@illinois.edu`