Correcting Multiple Substitutions in Nanopore-Sequencing Reads

Anisha Banerjee[1], **Yonatan Yehezkeally**[2], Antonia Wachter-Zeh[1], and Eitan Yaakobi[3]

[1]Technical University of Munich
Institute for Communications Engineering

[2]Newcastle University, School of Computing

[3]Technion – Israel Institute of Technology
Department of Computer Science

June 24, 2025

Funded by the European Union

Biotechnology and Biological Sciences Research Council

Outline

Motivation

- Need for dense, reliable, robust storage media
  - Molecular storage paradigms *e.g., DNA storage.*[1]

[1] R. Heckel *et al.*, "A characterization of the DNA data storage channel," *Scientific Reports*, 2019
[2] D. Deamer *et al.*, "Three decades of nanopore sequencing," *Nat. Biotech.*, 2016
[2] A. H. Laszlo *et al.*, "Decoding long nanopore sequencing reads of natural DNA," *Nat. Biotech.*, 2014

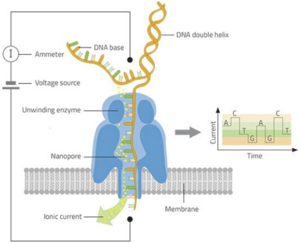
Motivation

- Need for dense, reliable, robust storage media
  - Molecular storage paradigms *e.g., DNA storage.*[1]
- Faster, cheaper sequencers in development.

[1] R. Heckel *et al.*, “A characterization of the DNA data storage channel,” *Scientific Reports*, 2019
[2] D. Deamer *et al.*, “Three decades of nanopore sequencing,” *Nat. Biotech.*, 2016
[2] A. H. Laszlo *et al.*, “Decoding long nanopore sequencing reads of natural DNA,” *Nat. Biotech.*, 2014

Motivation

- Need for dense, reliable, robust storage media
  - Molecular storage paradigms *e.g., DNA storage.*[1]
- Faster, cheaper sequencers in development.
- Nanopore sequencing[2] :
  - \+ Can read longer DNA strands
  - \+ More portable
  - – High error rates



Source: *"Decoding DNA with a pocket-sized sequencer," Science in School.* https://www.scienceinschool.org/article/2018/decoding-dna-pocket-sized-sequencer/

---

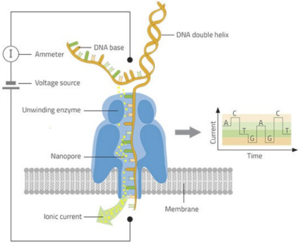
[1] R. Heckel *et al.*, "A characterization of the DNA data storage channel," *Scientific Reports*, 2019

[2] D. Deamer *et al.*, "Three decades of nanopore sequencing," *Nat. Biotech.*, 2016

[2] A. H. Laszlo *et al.*, "Decoding long nanopore sequencing reads of natural DNA," *Nat. Biotech.*, 2014

Motivation

- Need for dense, reliable, robust storage media
  - Molecular storage paradigms *e.g., DNA storage.*[1]
- Faster, cheaper sequencers in development.
- Nanopore sequencing[2] :
  - \+ Can read longer DNA strands
  - \+ More portable
  - − High error rates
- *Aim:* Design coding techniques tailored for nanopore sequencing!



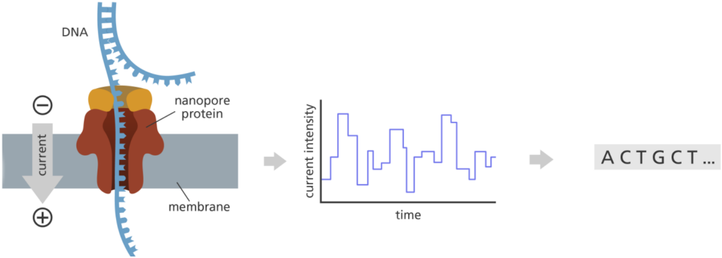
Source: *"Decoding DNA with a pocket-sized sequencer," Science in School.* `https://www.scienceinschool.org/article/2018/decoding-dna-pocket-sized-sequencer/`

[1] R. Heckel *et al.*, "A characterization of the DNA data storage channel," *Scientific Reports*, 2019

[2] D. Deamer *et al.*, "Three decades of nanopore sequencing," *Nat. Biotech.*, 2016

[2] A. H. Laszlo *et al.*, "Decoding long nanopore sequencing reads of natural DNA," *Nat. Biotech.*, 2014

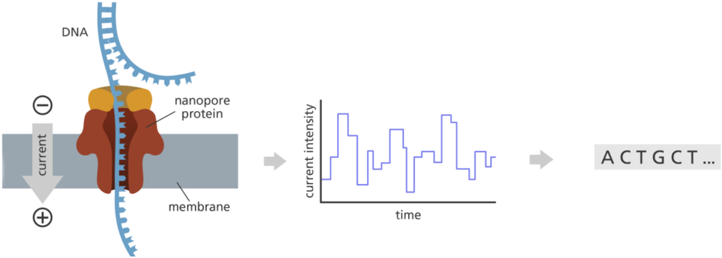
Nanopore Sequencing



Source: *"What is Oxford Nanopore Technology (ONT) sequencing?,"* https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/

[MDK18] W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018

Nanopore Sequencing



Source: *"What is Oxford Nanopore Technology (ONT) sequencing?,"* *https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/*

- Sources of noise [MDK18] :
  - Nanopore holds $\ell > 1$ nucleotides at a time $\rightarrow$ Intersymbol interference (ISI)!
  - Strand moves irregularly $\rightarrow$ backtracking & skipping.
  - Noisy measurements

[MDK18] W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018

Prior Work

- Channel model & understanding capacity: [MDK18] , [HCW21] , [MVS22] , [RW25]

[MDK18] W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018
[HCW21] R. Hulett *et al.*, "On coding for an abstracted nanopore channel for DNA storage," in *ISIT*, 2021
[MVS22] B. McBain *et al.*, "Finite-state semi-markov channels for nanopore sequencing," in *ISIT*, 2022
[RW25] A. Rameshwar and N. Weinberger, "On achievable rates over noisy nanopore channels," in *ISIT*, 2025
[CVVY21] Y. M. Chee *et al.*, "Coding for transverse-reads in domain wall memories," in *ISIT*, 2021

Prior Work

- Channel model & understanding capacity: [MDK18] , [HCW21] , [MVS22] , [RW25]
- Similar channel model for racetrack memories: [CVVY21]

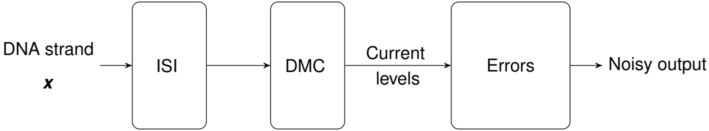
[MDK18] W. Mao *et al.*, “Models and information-theoretic bounds for nanopore sequencing,” *TIT*, 2018
[HCW21] R. Hulett *et al.*, “On coding for an abstracted nanopore channel for DNA storage,” in *ISIT*, 2021
[MVS22] B. McBain *et al.*, “Finite-state semi-markov channels for nanopore sequencing,” in *ISIT*, 2022
[RW25] A. Rameshwar and N. Weinberger, “On achievable rates over noisy nanopore channels,” in *ISIT*, 2025
[CVVY21] Y. M. Chee *et al.*, “Coding for transverse-reads in domain wall memories,” in *ISIT*, 2021

Prior Work

- Channel model & understanding capacity: [MDK18] , [HCW21] , [MVS22] , [RW25]
- Similar channel model for racetrack memories: [CVVY21]
- [MDK18] & [CVVY21] inspired a **simplified model** of nanopore sequencing.

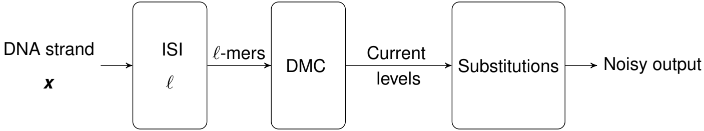
[MDK18] W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018
[HCW21] R. Hulett *et al.*, "On coding for an abstracted nanopore channel for DNA storage," in *ISIT*, 2021
[MVS22] B. McBain *et al.*, "Finite-state semi-markov channels for nanopore sequencing," in *ISIT*, 2022
[RW25] A. Rameshwar and N. Weinberger, "On achievable rates over noisy nanopore channels," in *ISIT*, 2025
[CVVY21] Y. M. Chee *et al.*, "Coding for transverse-reads in domain wall memories," in *ISIT*, 2021

Simplified Model



W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018

Simplified Model

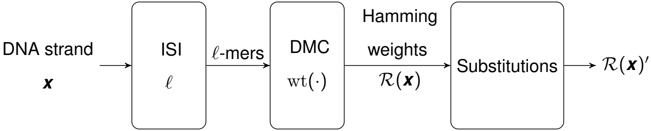


We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

W. Mao *et al.*, “Models and information-theoretic bounds for nanopore sequencing,” *TIT*, 2018

Simplified Model



We model

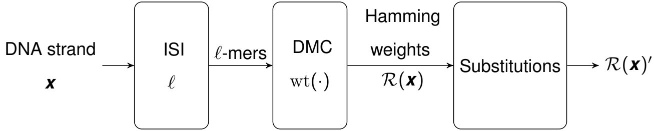
- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

Assumptions:

- Input $\boldsymbol{x} \in \{0, 1\}^n$
- DMC maps ℓ-mers to Hamming weights.

W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018

Simplified Model



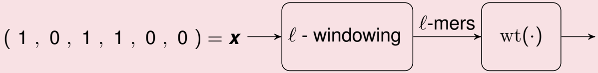
We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

Assumptions:

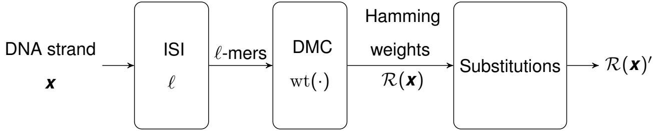
- Input $\boldsymbol{x} \in \{0,1\}^n$
- DMC maps ℓ-mers to Hamming weights.

Example ($\ell = 3$)



W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018

Simplified Model



We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

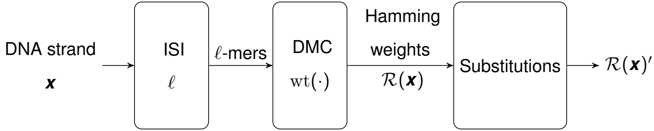
Assumptions:

- Input $\boldsymbol{x} \in \{0,1\}^n$
- DMC maps ℓ-mers to Hamming weights.

Example ($\ell = 3$)

W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018

Simplified Model



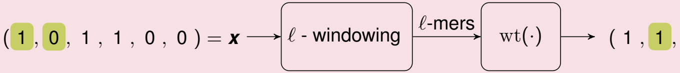
We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

Assumptions:

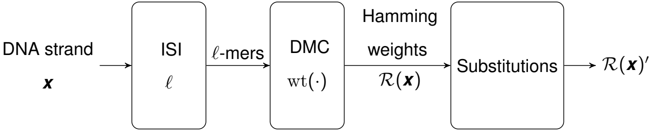
- Input $\boldsymbol{x} \in \{0, 1\}^n$
- DMC maps ℓ-mers to Hamming weights.

Example ($\ell = 3$)



W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018

Simplified Model



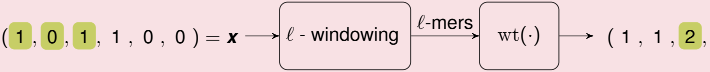
We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

Assumptions:

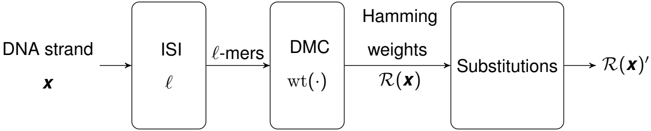
- Input $\boldsymbol{x} \in \{0,1\}^n$
- DMC maps ℓ-mers to Hamming weights.

Example ($\ell = 3$)



W. Mao *et al.*, “Models and information-theoretic bounds for nanopore sequencing,” *TIT*, 2018

Simplified Model



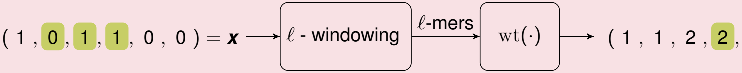
We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

Assumptions:

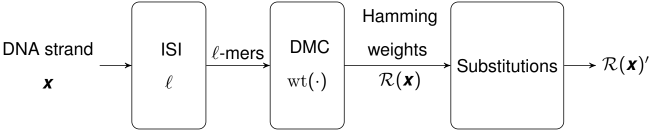
- Input $\boldsymbol{x} \in \{0,1\}^n$
- DMC maps ℓ-mers to Hamming weights.

Example ($\ell = 3$)



W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018

Simplified Model



We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

Assumptions:

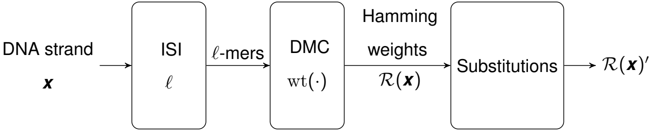
- Input $\boldsymbol{x} \in \{0,1\}^n$
- DMC maps ℓ-mers to Hamming weights.

Example ($\ell = 3$)

$$(1, 0, 1, 1, 0, 0) = \boldsymbol{x} \longrightarrow \boxed{\ell\text{ - windowing}} \xrightarrow{\ell\text{-mers}} \boxed{\mathrm{wt}(\cdot)} \longrightarrow (1, 1, 2, 2, 2,$$

W. Mao *et al.*, “Models and information-theoretic bounds for nanopore sequencing,” *TIT*, 2018

Simplified Model



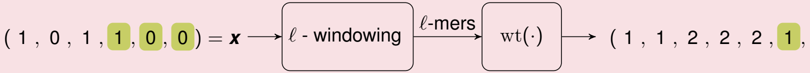
We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

Assumptions:

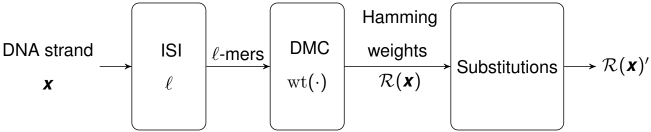
- Input $\boldsymbol{x} \in \{0, 1\}^n$
- DMC maps ℓ-mers to Hamming weights.

Example ($\ell = 3$)



W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018

Simplified Model



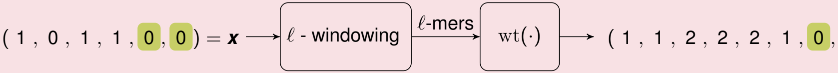
We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

Assumptions:

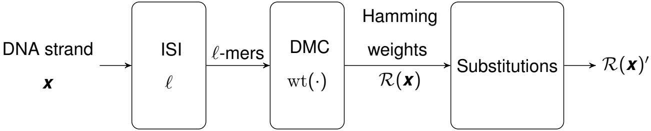
- Input $\boldsymbol{x} \in \{0, 1\}^n$
- DMC maps ℓ-mers to Hamming weights.

Example ($\ell = 3$)



W. Mao *et al.*, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, 2018

Simplified Model



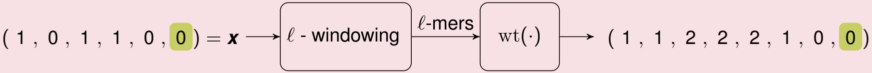
We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

Assumptions:

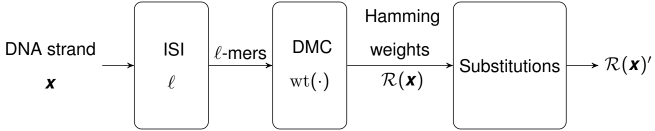
- Input $\boldsymbol{x} \in \{0, 1\}^n$
- DMC maps ℓ-mers to Hamming weights.

Example ($\ell = 3$)



W. Mao *et al.*, “Models and information-theoretic bounds for nanopore sequencing,” *TIT*, 2018

Simplified Model



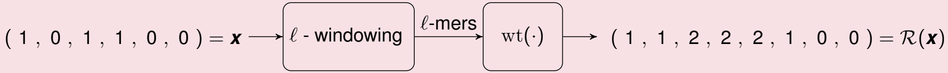
We model

- ISI as length-ℓ-sliding window
- Measurement noise as substitutions

Assumptions:

- Input $\boldsymbol{x} \in \{0,1\}^n$
- DMC maps ℓ-mers to Hamming weights.

Example ($\ell = 3$)



W. Mao *et al.*, “Models and information-theoretic bounds for nanopore sequencing,” *TIT*, 2018

Simplified Model

Definition

The ℓ-***read vector*** of any $\boldsymbol{x} \in \Sigma_2^n$ is

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n))$$

where $\mathrm{wt}(\cdot)$ denotes Hamming weight.

Y. M. Chee *et al.*, "Coding for transverse-reads in domain wall memories," in *ISIT*, 2021

Simplified Model

Definition

The ℓ-***read vector*** of any $\boldsymbol{x} \in \Sigma_2^n$ is

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n))$$

where $\mathrm{wt}(\cdot)$ denotes Hamming weight.

- Read vectors have special properties.

Y. M. Chee *et al.*, “Coding for transverse-reads in domain wall memories,” in *ISIT*, 2021

Simplified Model

Definition

The ℓ-***read vector*** of any $\boldsymbol{x} \in \Sigma_2^n$ is

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n))$$

where $\mathrm{wt}(\cdot)$ denotes Hamming weight.

- Read vectors have special properties.
  - $|\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1}| \leq 1 \impliedby \mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$

Y. M. Chee *et al.*, "Coding for transverse-reads in domain wall memories," in *ISIT*, 2021

Simplified Model

Definition

The ℓ-***read vector*** of any $\boldsymbol{x} \in \Sigma_2^n$ is

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n))$$

where $\mathrm{wt}(\cdot)$ denotes Hamming weight.

- Read vectors have special properties.
  - $|\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1}| \leq 1$,
  - $\sum_i \mathcal{R}(\boldsymbol{x})_i \bmod \ell = 0 \Longleftarrow \sum_i \mathcal{R}(\boldsymbol{x})_i = \ell \cdot \mathrm{wt}(\boldsymbol{x})$

Y. M. Chee *et al.*, “Coding for transverse-reads in domain wall memories,” in *ISIT*, 2021

Simplified Model

Definition

The ℓ-***read vector*** of any $\boldsymbol{x} \in \Sigma_2^n$ is

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n))$$

where $\mathrm{wt}(\cdot)$ denotes Hamming weight.

- Read vectors have special properties.
  - $|\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1}| \leq 1$,
  - $\sum_i \mathcal{R}(\boldsymbol{x})_i \bmod \ell = 0$, etc.
- Can correct substitution errors of abs. value ≥ 2.

Y. M. Chee *et al.*, “Coding for transverse-reads in domain wall memories,” in *ISIT*, 2021

Prior Work

- Error correction for the *read vector* channel
  - [BYWY24] , [BYWY25] , [CIV24] , [SG25] , [YYS25] ....
- *Weighted Read Vector channel* [YEY24]

[BYWY24] A. Banerjee *et al.*, "Error-correcting codes for nanopore sequencing," *TIT*, 2024

[BYWY25] A. Banerjee *et al.*, "Correcting a single deletion in reads from a nanopore sequencer," in *ISIT*, 2024

[CIV24] Y. M. Chee *et al.*, "Coding scheme for noisy nanopore sequencing with backtracking and skipping errors," in *ISIT*, 2024

[SG25] Y. Sun and G. Ge, "Bounds and constructions of ℓ-read codes under the hamming metric," *TIT*, 2025

[YYS25] W. Yu *et al.*, "On the asymptotic rate of optimal codes that correct tandem duplications for nanopore sequencing," *TIT*, 2025

[YEY24] O. Yerushalmi *et al.*, "The capacity of the weighted read channel," in *ISIT*, 2024

ℓ-Read Codes

- Read vectors possess inherent redundancy.
  - May aid in error correction!

ℓ-Read Codes

- Read vectors possess inherent redundancy.
  - May aid in error correction!

Definition [Error-correcting Code]

$\mathcal{C}$ is an ℓ-**read code** if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$,

$$E(\mathcal{R}(\boldsymbol{x})) \cap E(\mathcal{R}(\boldsymbol{y})) = \emptyset,$$

where $E(\cdot)$ is the error ball.

Comparison with the Classical Channel

- Do ℓ-**read codes** require lesser redundancy?

[15] A. Banerjee *et al.*, "Error-correcting codes for nanopore sequencing," *TIT*, 2024
[16] A. Banerjee *et al.*, "Correcting a single deletion in reads from a nanopore sequencer," in *ISIT*, 2024

Comparison with the Classical Channel

- Do ℓ-**read codes** require lesser redundancy?
  - ‘Classical’ $\rightarrow \ell = 1$.
  - ‘Nanopore’ $\rightarrow \ell \geq 2$.

[15] A. Banerjee *et al.*, “Error-correcting codes for nanopore sequencing,” *TIT*, 2024

[16] A. Banerjee *et al.*, “Correcting a single deletion in reads from a nanopore sequencer,” in *ISIT*, 2024

Comparison with the Classical Channel

- Do ℓ-**read codes** require lesser redundancy?
  - 'Classical' $\to \ell = 1$.
  - 'Nanopore' $\to \ell \geq 2$.

	Classical	Nanopore
1 substitution[15]	$\log n$	$\log \log n$
1 deletion[16]	$\log n$	$\log n - \ell$
1 deletion, 2 reads[16]	$\log \log n$	0

[15] A. Banerjee *et al.*, "Error-correcting codes for nanopore sequencing," *TIT*, 2024

[16] A. Banerjee *et al.*, "Correcting a single deletion in reads from a nanopore sequencer," in *ISIT*, 2024

Comparison with the Classical Channel

- Do ℓ-**read codes** require lesser redundancy?
  - ‘Classical’ $\rightarrow \ell = 1$.
  - ‘Nanopore’ $\rightarrow \ell \geq 2$.

	Classical	**Nanopore**
1 substitution[15]	$\log n$	$\log \log n$
1 deletion[16]	$\log n$	$\log n - \ell$
1 deletion, 2 reads[16]	$\log \log n$	0
$t \geq 2$ substitutions	$t \log n$	?

[15] A. Banerjee *et al.*, “Error-correcting codes for nanopore sequencing,” *TIT*, 2024

[16] A. Banerjee *et al.*, “Correcting a single deletion in reads from a nanopore sequencer,” in *ISIT*, 2024

Comparison with the Classical Channel

- Do ℓ-**read codes** require lesser redundancy?
  - 'Classical' $\rightarrow \ell = 1$.
  - 'Nanopore' $\rightarrow \ell \geq 2$.

	Classical	**Nanopore**	
1 substitution[15]	$\log n$	$\log \log n$	
1 deletion[16]	$\log n$	$\log n - \ell$	😢
1 deletion, 2 reads[16]	$\log \log n$	0	😁
$t \geq 2$ substitutions	$t \log n$	$t \log n$	😭

[15] A. Banerjee *et al.*, "Error-correcting codes for nanopore sequencing," *TIT*, 2024

[16] A. Banerjee *et al.*, "Correcting a single deletion in reads from a nanopore sequencer," in *ISIT*, 2024

Minimum Redundancy

Theorem

The redundancy of any t**-substitution** ℓ**-read code**, for $t, \ell \geq 2$, is bounded from below by

$$t \log n - O(1).$$

Minimum Redundancy

Theorem

The redundancy of any ***t*-substitution ℓ-read code**, for $t, \ell \geq 2$, is bounded from below by

$$t \log n - O(1).$$

- But for $t = 1$, we need $\log \log n - o(1)$ redundant bits.

Minimum Redundancy

Theorem

The redundancy of any t**-substitution** ℓ**-read code**, for $t, \ell \geq 2$, is bounded from below by

$$t \log n - O(1).$$

- But for $t = 1$, we need $\log \log n - o(1)$ redundant bits.

Why? 🥲

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^{i})$ and $x_i = 0$ for $i \notin [1, n]$.

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^{i})$ and $x_i = 0$ for $i \notin [1, n]$.

- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ 3\ ,\ 2\ ,\ 1\ ,\ 0\ ,\ 0\ )$$

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^{i})$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\,1\,,\ 1\ ,\ 2\ ,\ 3\ ,\ 2\ ,\ 1\ ,\ 0\ ,\ 0\ )$$

Left to right reconstruction:

- $x_1' = 1$,

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^{i})$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ 3\ ,\ 2\ ,\ 1\ ,\ 0\ ,\ 0\ )$$

Left to right reconstruction:

- $x_1' = 1$, $x_2' = 0$,

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^{i})$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\,1\,,\,1\,,\,\boxed{2}\,,\,3\,,\,2\,,\,1\,,\,0\,,\,0\,)$$

Left to right reconstruction:

- $x_1' = 1$, $x_2' = 0$, $x_3' = 1$,

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}^{i}_{i-\ell+1})$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ \boxed{3}\ ,\ 2\ ,\ 1\ ,\ 0\ ,\ 0\ )$$

Left to right reconstruction:

- $x'_1 = 1$, $x'_2 = 0$, $x'_3 = 1$, $x'_4 = 2$ ✗

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ 3\ ,\ 2\ ,\ 1\ ,\ 0\ ,\ 0\ )$$

Left to right reconstruction:

- $x'_1 = 1$, $x'_2 = 0$, $x'_3 = 1$, $x'_4 = 2$ ✗
- Error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_4)$.

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ 3\ ,\ 2\ ,\ 1\ ,\ 0\ ,\ \boxed{0}\ )$$

Left to right reconstruction:

- $x_1' = 1$, $x_2' = 0$, $x_3' = 1$, $x_4' = 2$ ✗
- Error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_4)$.

Right to left reconstruction:

$x_6' = 0$

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ 3\ ,\ 2\ ,\ 1\ ,\ \boxed{0}\ ,\ 0\ )$$

Left to right reconstruction:

- $x_1' = 1$, $x_2' = 0$, $x_3' = 1$, $x_4' = 2$ ✗
- Error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_4)$.

Right to left reconstruction:

$x_5' = 0$, $x_6' = 0$

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^{i})$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ 3\ ,\ 2\ ,\ \boxed{1}\ ,\ 0\ ,\ 0\ )$$

Left to right reconstruction:

- $x'_1 = 1$, $x'_2 = 0$, $x'_3 = 1$, $x'_4 = 2$ ✗
- Error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_4)$.

Right to left reconstruction:

$x'_4 = 1$, $x'_5 = 0$, $x'_6 = 0$

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}^i_{i-\ell+1})$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ 3\ ,\ \boxed{2}\ ,\ 1\ ,\ 0\ ,\ 0\ )$$

Left to right reconstruction:

- $x'_1 = 1$, $x'_2 = 0$, $x'_3 = 1$, $x'_4 = 2$ ✗
- Error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_4)$.

Right to left reconstruction:

$x'_3 = 1$, $x'_4 = 1$, $x'_5 = 0$, $x'_6 = 0$

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ 3\ ,\ 2\ ,\ 1\ ,\ 0\ ,\ 0\ )$$

Left to right reconstruction:

- $x_1' = 1$, $x_2' = 0$, $x_3' = 1$, $x_4' = 2$ ✗
- Error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_4)$.

Right to left reconstruction:

$x_2' = 1$, $x_3' = 1$, $x_4' = 1$, $x_5' = 0$, $x_6' = 0$

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ \boxed{2},\ 3\ ,\ 2\ ,\ 1\ ,\ 0\ ,\ 0\ )$$

Left to right reconstruction:

- $x_1' = 1$, $x_2' = 0$, $x_3' = 1$, $x_4' = 2$ ✗
- Error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_4)$.

Right to left reconstruction:

- $x_1' = 0$, $x_2' = 1$, $x_3' = 1$, $x_4' = 1$, $x_5' = 0$, $x_6' = 0$

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^{i})$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{\text{1substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ 3\ ,\ 2\ ,\ 1\ ,\ 0\ ,\ 0\ )$$

Left to right reconstruction:

- $x_1' = 1$, $x_2' = 0$, $x_3' = 1$, $x_4' = 2$ ✗
- Error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_4)$.

Right to left reconstruction:

- $x_1' = 0$, $x_2' = 1$, $x_3' = 1$, $x_4' = 1$, $x_5' = 0$, $x_6' = 0$
- No inconsistency, error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_8)$.

Reconstruction from Noisy Read Vector

- Recall that for any $\boldsymbol{x} \in \Sigma_2^n$,
  - $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^{i})$ and $x_i = 0$ for $i \notin [1, n]$.
- Can reconstruct $\boldsymbol{x}$ from $\mathcal{R}(\boldsymbol{x})$ with
  - $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Example ($\ell = 3$)

$$\mathcal{R}(\boldsymbol{x}) \xrightarrow{1\text{substitution}} \mathcal{R}(\boldsymbol{x})' = (\ 1\ ,\ 1\ ,\ 2\ ,\ 3\ ,\ 2\ ,\ 1\ ,\ 0\ ,\ 0\ )$$

Left to right reconstruction:

- $x_1' = 1$, $x_2' = 0$, $x_3' = 1$, $x_4' = 2$ ✗
- Error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_4)$.

Right to left reconstruction:

- $x_1' = 0$, $x_2' = 1$, $x_3' = 1$, $x_4' = 1$, $x_5' = 0$, $x_6' = 0$
- No inconsistency, error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_8)$.

Error in $(\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_4)$.

Reconstruction from Noisy Read Vector

- Recall that
  - $\sum_i \mathcal{R}(\boldsymbol{x})_i \bmod \ell = 0$
  - Only ± 1 errors are non-trivial to correct.

Reconstruction from Noisy Read Vector

- Recall that
  - $\sum_i \mathcal{R}(\boldsymbol{x})_i \bmod \ell = 0$
  - Only ± 1 errors are non-trivial to correct.
- Thus 1 substitution is detectable

Reconstruction from Noisy Read Vector

- Recall that
  - $\sum_i \mathcal{R}(\boldsymbol{x})_i \bmod \ell = 0$
  - Only ± 1 errors are non-trivial to correct.
- Thus 1 substitution is detectable & localizable!

Reconstruction from Noisy Read Vector

- Recall that
  - $\sum_i \mathcal{R}(\boldsymbol{x})_i \bmod \ell = 0$
  - Only ± 1 errors are non-trivial to correct.
- Thus 1 substitution is detectable & localizable!
  - *But not under multiple substitutions!*

Reconstruction from Noisy Read Vector

- Recall that
  - $\sum_i \mathcal{R}(\boldsymbol{x})_i \bmod \ell = 0$
  - Only ± 1 errors are non-trivial to correct.
- Thus 1 substitution is detectable & localizable!
  - *But not under multiple substitutions!*

Example ($\ell = 3$)

$$\boldsymbol{x} = (\,1\,,\,0\,,\,1\,,\,1\,,\,0\,,\,0\,),\;\; \mathcal{R}(\boldsymbol{x}) = (\,1\,,\,1\,,\,2\,,\,2\,,\,2\,,\,1\,,\,0\,,\,0\,)$$

$$\boldsymbol{y} = (\,0\,,\,1\,,\,1\,,\,0\,,\,1\,,\,0\,),\;\; \mathcal{R}(\boldsymbol{y}) = (\,0\,,\,1\,,\,2\,,\,2\,,\,2\,,\,1\,,\,1\,,\,0\,)$$

$$d_H(\mathcal{R}(\boldsymbol{x}), \mathcal{R}(\boldsymbol{y})) = 2$$

Reconstruction from Noisy Read Vector

- Recall that
  - $\sum_i \mathcal{R}(\boldsymbol{x})_i \bmod \ell = 0$
  - Only ± 1 errors are non-trivial to correct.
- Thus 1 substitution is detectable & localizable!
  - *But not under multiple substitutions!*

Example ($\ell = 3$)

$$\boldsymbol{x} = (\boxed{1}, \boxed{0}, 1, \boxed{1}, \boxed{0}, 0), \quad \mathcal{R}(\boldsymbol{x}) = (\boxed{1}, 1, 2, 2, 2, 1, \boxed{0}, 0)$$

$$\boldsymbol{y} = (\boxed{0}, \boxed{1}, 1, \boxed{0}, \boxed{1}, 0), \quad \mathcal{R}(\boldsymbol{y}) = (\boxed{0}, 1, 2, 2, 2, 1, \boxed{1}, 0)$$

$$d_H(\mathcal{R}(\boldsymbol{x}), \mathcal{R}(\boldsymbol{y})) = 2$$

Reconstruction from Noisy Read Vector

- Recall that
  - $\sum_i \mathcal{R}(\boldsymbol{x})_i \bmod \ell = 0$
  - Only ± 1 errors are non-trivial to correct.
- Thus 1 substitution is detectable & localizable!
  - *But not under multiple substitutions!*

Example ($\ell = 3$)

$$\boldsymbol{x} = (1, 0, 1, 1, 0, 0), \quad \mathcal{R}(\boldsymbol{x}) = (1, 1, 2, 2, 2, 1, 0, 0)$$

$$\boldsymbol{y} = (0, 1, 1, 0, 1, 0), \quad \mathcal{R}(\boldsymbol{y}) = (0, 1, 2, 2, 2, 1, 1, 0)$$

$$d_H(\mathcal{R}(\boldsymbol{x}), \mathcal{R}(\boldsymbol{y})) = 2$$

- 2 substitutions can lead to a valid ℓ-read vector

Reconstruction from Noisy Read Vector

- Recall that
  - $\sum_i \mathcal{R}(\boldsymbol{x})_i \bmod \ell = 0$
  - Only ± 1 errors are non-trivial to correct.
- Thus 1 substitution is detectable & localizable!
  - *But not under multiple substitutions!*

Example ($\ell = 3$)

$$\boldsymbol{x} = (\,1\,,\,0\,,\,1\,,\,1\,,\,0\,,\,0\,),\ \ \mathcal{R}(\boldsymbol{x}) = (\,1\,,\,1\,,\,2\,,\,2\,,\,2\,,\,1\,,\,0\,,\,0\,)$$

$$\boldsymbol{y} = (\,0\,,\,1\,,\,1\,,\,0\,,\,1\,,\,0\,),\ \ \mathcal{R}(\boldsymbol{y}) = (\,0\,,\,1\,,\,2\,,\,2\,,\,2\,,\,1\,,\,1\,,\,0\,)$$

$$d_H(\mathcal{R}(\boldsymbol{x}), \mathcal{R}(\boldsymbol{y})) = 2$$

- 2 substitutions can lead to a valid ℓ-read vector $\implies$ can't localize!

Naive Construction

Definition (Read vector)

The ℓ-*read vector* of any $\boldsymbol{x} \in \Sigma_2^n$ is of length $n+\ell-1$ and denoted by

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n)).$$

Here $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ where for $j \notin [n]$, let $x_j = 0$.

Naive Construction

Definition (Read vector)

The ℓ-*read vector* of any $\boldsymbol{x} \in \Sigma_2^n$ is of length $n+\ell-1$ and denoted by

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n)).$$

Here $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ where for $j \notin [n]$, let $x_j = 0$.

- $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.

Naive Construction

Definition (Read vector)

The ℓ-*read vector* of any $\boldsymbol{x} \in \Sigma_2^n$ is of length $n + \ell - 1$ and denoted by

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n)).$$

Here $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ where for $j \notin [n]$, let $x_j = 0$.

- $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.
- $\boldsymbol{x}$ uniquely determined from n-prefix of $\mathcal{R}(\boldsymbol{x})$.

Naive Construction

Definition (Read vector)

The ℓ-*read vector* of any $\boldsymbol{x} \in \Sigma_2^n$ is of length $n + \ell - 1$ and denoted by

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n)).$$

Here $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ where for $j \notin [n]$, let $x_j = 0$.

- $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.
- $\boldsymbol{x}$ uniquely determined from n-prefix of $\mathcal{R}(\boldsymbol{x})$ mod 2

Naive Construction

Definition (Read vector)

The ℓ-*read vector* of any $\boldsymbol{x} \in \Sigma_2^n$ is of length $n + \ell - 1$ and denoted by

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n)).$$

Here $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ where for $j \notin [n]$, let $x_j = 0$.

- $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.
- $\boldsymbol{x}$ uniquely determined from n-prefix of $\mathcal{R}(\boldsymbol{x})$ mod 2 $\leftarrow$ Use any t-substitution-correcting code!

Naive Construction

Definition (Read vector)

The ℓ-*read vector* of any $\boldsymbol{x} \in \Sigma_2^n$ is of length $n+\ell-1$ and denoted by

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \ldots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n)).$$

Here $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ where for $j \notin [n]$, let $x_j = 0$.

- $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.
- $\boldsymbol{x}$ uniquely determined from n-prefix of $\mathcal{R}(\boldsymbol{x})$ mod 2 $\leftarrow$ Use any t-substitution-correcting code!

Construction (t-substitution-ℓ-read code)

$$\{\boldsymbol{x} \in \Sigma_2^n : (\mathcal{R}(\boldsymbol{x})_1, \ldots, \mathcal{R}(\boldsymbol{x})_n) \bmod 2 \in \mathcal{C}(n,t)\},$$

where $\mathcal{C}(n,t) \subset \Sigma_2^n$ is a t-substitution-correcting code.

Naive Construction

Definition (Read vector)

The ℓ-*read vector* of any $\boldsymbol{x} \in \Sigma_2^n$ is of length $n+\ell-1$ and denoted by

$$\mathcal{R}(\boldsymbol{x}) = (\mathrm{wt}(x_1), \mathrm{wt}(\boldsymbol{x}_1^2), \dots, \mathrm{wt}(\boldsymbol{x}_{n-1}^n), \mathrm{wt}(x_n)).$$

Here $\mathcal{R}(\boldsymbol{x})_i = \mathrm{wt}(\boldsymbol{x}_{i-\ell+1}^i)$ where for $j \notin [n]$, let $x_j = 0$.

- $\mathcal{R}(\boldsymbol{x})_i - \mathcal{R}(\boldsymbol{x})_{i-1} = x_i - x_{i-\ell}$.
- $\boldsymbol{x}$ uniquely determined from n-prefix of $\mathcal{R}(\boldsymbol{x})$ mod 2 $\leftarrow$ Use any t-substitution-correcting code!

Construction (t-substitution-ℓ-read code)

$$\{\boldsymbol{x} \in \Sigma_2^n : (\mathcal{R}(\boldsymbol{x})_1, \dots, \mathcal{R}(\boldsymbol{x})_n) \bmod 2 \in \mathcal{C}(n,t)\},$$

where $\mathcal{C}(n,t) \subset \Sigma_2^n$ is a t-substitution-correcting code.

$\rightarrow$ Optimal up to additive constant.

Summary

Results

For a simplified model of nanopore sequencing, we show

- $t \log n - O(1)$ min redundancy needed to correct $t \geq 2$ substitutions.
- Naive construction, optimal up to additive constant.

Summary

Results

For a simplified model of nanopore sequencing, we show

- $t \log n - O(1)$ min redundancy needed to correct $t \geq 2$ substitutions.
- Naive construction, optimal up to additive constant.

Future work

- Multiple deletions & combination with substitutions.
- Weighted read vectors

Summary

Results

For a simplified model of nanopore sequencing, we show

- $t \log n - O(1)$ min redundancy needed to correct $t \geq 2$ substitutions.
- Naive construction, optimal up to additive constant.

Future work

- Multiple deletions & combination with substitutions.
- Weighted read vectors

Thank you!

References I

[1] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Scientific Reports*, vol. 9, no. 9663, Jul. 2019.

[2] D. Deamer, M. Akeson, and D. Branton, "Three decades of nanopore sequencing," *Nat. Biotech.*, vol. 34, no. 5, pp. 518–524, May 2016.

[3] A. H. Laszlo *et al.*, "Decoding long nanopore sequencing reads of natural DNA," *Nat. Biotech.*, vol. 32, no. 8, pp. 829–833, Aug. 2014.

[4] W. Mao, S. N. Diggavi, and S. Kannan, "Models and information-theoretic bounds for nanopore sequencing," *TIT*, vol. 64, no. 4, pp. 3216–3236, Apr. 2018.

[5] R. Hulett, S. Chandak, and M. Wootters, "On coding for an abstracted nanopore channel for DNA storage," in *ISIT*, Jul. 2021, pp. 2465–2470.

[6] B. McBain, E. Viterbo, and J. Saunderson, "Finite-state semi-markov channels for nanopore sequencing," in *ISIT*, Jun. 2022, pp. 216–221.

References II

[7] A. Rameshwar and N. Weinberger, “On achievable rates over noisy nanopore channels,” in *ISIT*, 2025, (arXiv preprint arXiv:2501.02917).

[8] Y. M. Chee, A. Vardy, V. K. Vu, and E. Yaakobi, “Coding for transverse-reads in domain wall memories,” in *ISIT*, Jul. 2021, pp. 2924–2929.

[9] A. Banerjee, Y. Yehezkeally, A. Wachter-Zeh, and E. Yaakobi, “Error-correcting codes for nanopore sequencing,” *TIT*, 2024.

[10] A. Banerjee, Y. Yehezkeally, A. Wachter-Zeh, and E. Yaakobi, “Correcting a single deletion in reads from a nanopore sequencer,” in *ISIT*, Jul. 2024, pp. 103–108.

[11] Y. M. Chee, K. A. Schouhamer Immink, and V. K. Vu, “Coding scheme for noisy nanopore sequencing with backtracking and skipping errors,” in *ISIT*, 2024.

[12] Y. Sun and G. Ge, “Bounds and constructions of ℓ-read codes under the hamming metric,” *TIT*, pp. 1–1, 2025.

[13] W. Yu, Z. Ye, and M. Schwartz, “On the asymptotic rate of optimal codes that correct tandem duplications for nanopore sequencing,” *TIT*, vol. 71, no. 5, pp. 3569–3581, May 2025.

References III

[14] O. Yerushalmi, T. Etzion, and E. Yaakobi, "The capacity of the weighted read channel," in *ISIT*, 2024, (arXiv preprint arXiv:2401.15368).