On the Capacity of DNA-based Data Storage under Substitution Errors

Andreas Lenz\textsuperscript{1}, Paul H. Siegel\textsuperscript{2}, Antonia Wachter-Zeh\textsuperscript{1}, Eitan Yaakobi\textsuperscript{3}

\textsuperscript{1}Institute for Communications Engineering  
Technical University of Munich

\textsuperscript{2}Department of Electrical and Computer Engineering  
University of California

\textsuperscript{3}Computer Science Department, Technion  
Israel Institute of Technology

May 9th, 2022
Outline

- Channel Model
- Related Work
- Preliminaries
- Channel Capacity
- Summary & Outlook
Data Storage in DNA

High density data storage
- DNA: $10^9$ GB/mm$^3$
- Tape: 10 – 100 GB/mm$^3$

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Data Storage in DNA

- **Never obsolete**
- **High density data storage**
  - **DNA:** $10^9$ GB/mm$^3$
  - **Tape:** $10 - 100$ GB/mm$^3$
Data Storage in DNA

- **High density data storage**
  - DNA: $10^9$ GB/mm$^3$
  - Tape: $10 - 100$ GB/mm$^3$

- **Long term data storage**
  - (DNA from mammoths)

- **Never obsolete**
Data Storage in DNA

- High density data storage
  - DNA: $10^9$ GB/mm$^3$
  - Tape: 10 – 100 GB/mm$^3$

- Never obsolete

- Long term data storage (DNA from mammoths)

- Easily duplicatable (PCR)

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Data Storage in DNA

High density data storage

- DNA: $10^9$ GB/mm$^3$
- Tape: $10^{-100}$ GB/mm$^3$

Never obsolete

Easily duplicatable (PCR)

Cost per Genome

Long term data storage (DNA from mammoths)

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Data Storage in DNA

User Binary Data
000001101011001
110100010010101
101000111110100
Data Storage in DNA

User Binary Data
000001101011001
110100010010101
10100011110100

Encoding

DNA strands
GCTATGAGTACT
ATGATTGACTCT
GATGGCATAGCT

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Data Storage in DNA

User Binary Data
000001101011001
110100010010101
101000111110100

Encoding

DNA strands
GCTATGAGTACT
ATGATTGACTCT
GATGGCATAGCT

DNA Sequencer

DNA Synthesizer

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Data Storage in DNA

User Binary Data
000001101011001
110100010010101
101000111110100

Encoding

DNA strands
GCTATGAGTACT
ATGATTGACTCT
GATGGCATAGCT

DNA Synthesizer

Storage Container

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Data Storage in DNA

User Binary Data
000001101011001
110100010010101
101000111110100

DNA Synthesizer
DNA Sequencer

Storage Container

Encoding

DNA strands
GCTATGAGTACT
ATGATTGACTCT
GATGGCATAGCT

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Data Storage in DNA

User Binary Data:
000001101011001110100010010101101000111101001010001111010100

DNA Synthesizer

DNA strands
ATAATTGAGTCT
GCTGGCATAGCT
GATAGCTTAGCT
ATGATTGACTCT
GATGGCATACCT

DNA Sequencer

DNA Synthesizer

DNA strands
GCTATGAGTACT
ATGATTGACTCT
GATGGCATAGCT

DNA Sequencer

Storage Container

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Model

\[
\begin{align*}
X & \quad \text{Draw & Distort} \quad Y \\
X_1 & \quad \text{GCTATGAGTACT} \\
X_2 & \quad \text{ATGATTGACTCT} \\
X_3 & \quad \text{GATGGCATAGCT}
\end{align*}
\]

\[j = X_i + E_j:\]
- \(i.i.d.\) uniform random draws
- \(E_j: \) random error vectors (error probability \(p\))

In this work: Quaternary sequences (\(\mathbb{Z}_4 = \{A, C, G, T\}\))
Channel Model

\[ X \xrightarrow{\text{Draw & Distort}} Y \]

\( X_1 \quad \text{GCTATGAGTACT} \)
\( X_2 \quad \text{ATGATTGACTCT} \)
\( X_3 \quad \text{GATGGCATAGCT} \)

\( Y \quad \text{ATAATTGAGTCT} \)

\( Y_1 \)

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Model

\[ X \xrightarrow{\text{Draw \\ & Distort}} Y \]

\[ X_1 \rightarrow \text{GCTATGAGTACT} \rightarrow \text{ATAATTGAGTCT} \]
\[ X_2 \rightarrow \text{ATGATTGACTCT} \rightarrow \text{GCTGGCATAGCT} \]
\[ X_3 \rightarrow \text{GATGGCATACCT} \rightarrow \text{GCTGGCATAGCT} \]

\[ Y_1 \]
\[ Y_2 \]
Channel Model

\[ X \xrightarrow{\text{Draw & Distort}} Y \]

\begin{align*}
X_1 & \quad \text{GCTATGAGTACT} \quad \text{GCTGGCATAGCT} \quad \text{Y}_1 \\
X_2 & \quad \text{ATGATTGACTCT} \quad \text{GATAGCTTAGCT} \quad \text{Y}_2 \\
X_3 & \quad \text{GATGGCATAGCT} \quad \text{Y}_3
\end{align*}

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Model

\[ X \xrightarrow{\text{Draw & Distort}} Y \]

- \( X_1: \) GCTATGAGTACT
- \( X_2: \) ATGATTGACTCT
- \( X_3: \) GATGGCATAGCT

\( Y_1: \) ATAATTGAGTCT
\( Y_2: \) GCTGGGCATAGCT
\( Y_3: \) GATAGCTTAGCT
\( Y_4: \) ATGATTGACTCT

Ref: Lenz, Siegel, Wachter-Zeh, Yaakobi. "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Model

Draw & Distort

\[ X \quad \text{Draw & Distort} \quad Y \]

\[
\begin{align*}
X_1 & : \text{GCTATGAGTACT} & \rightarrow & : \text{ATAATTGAGTCT} \\
X_2 & : \text{ATGATTGACTCT} & \rightarrow & : \text{GCTGGCATAGCT} \\
X_3 & : \text{GATGGCATAAGCT} & \rightarrow & : \text{GATAGCTTAGCT} \\
& & & \rightarrow \text{Y_3} \\
& & & \rightarrow \text{Y_4} \\
& & & \rightarrow \text{Y_5} \\
& & & \rightarrow \text{Y_1} \\
& & & \rightarrow \text{Y_2} \\
& & & \rightarrow \text{Y_3} \\
& & & \rightarrow \text{Y_4} \\
& & & \rightarrow \text{Y_5} \\
\end{align*}
\]
Channel Model

\[ Y_j = X_{I_j} + E_j \]

- \( I_j \): i.i.d. uniform random draws
- \( E_j \): random error vectors (error probability \( p \))

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Model

\[
Y_j = X_{I_j} + E_j
\]

- \( I_j \): i.i.d. uniform random draws
- \( E_j \): random error vectors (error probability \( p \))
- In this work: Quaternary sequences (\( \mathbb{Z}_4 = \{ A, C, G, T \} \))
Channel Model

Channel Input:
- $M$ Sequences, each of length $L$
- $X = (X_1, \ldots, X_M)$
Channel Model

Channel Input:
- $M$ Sequences, each of length $L$
- $X = (X_1, \ldots, X_M)$
- $\beta = \log_4 \frac{M}{L}$

![Diagram showing the channel model with DNA sequences and their corresponding outputs.](diagram.jpg)
Channel Input:
- $M$ Sequences, each of length $L$
- $X = (X_1, \ldots, X_M)$
- $\beta = \log_4 \frac{M}{L}$

Channel Output:
- $N$ sequences, each of length $L$
- $Y = (Y_1, \ldots, Y_N)$
Channel Model

\[ X \quad \text{Draw & Distort} \quad Y \]

\[
\begin{align*}
X_1 & \quad \text{GCTATGAGTACT} & Y_1 & \quad \text{ATAATTGAGTCT} \\
X_2 & \quad \text{ATGATTGACTCT} & Y_2 & \quad \text{GCTGGCATAGCT} \\
X_3 & \quad \text{GATGGCATAGCT} & Y_3 & \quad \text{GATAGCTTAGCT} \\
\end{align*}
\]

Channel Input:
- \( M \) Sequences, each of length \( L \)
- \( X = (X_1, \ldots, X_M) \)
- \( \beta = \log_4 \frac{M}{L} \)

Channel Output:
- \( N \) sequences, each of length \( L \)
- \( Y = (Y_1, \ldots, Y_N) \)
- \( c = \frac{N}{M} \)
Channel Model - Codes and Information Rates

Communication System:

Message

\[ W \]
Channel Model - Codes and Information Rates

Communication System:

Message

\( W \) ← Encoder

\[ C = \{ X(1), \ldots, X(4) \} \subset \mathbb{Z}^{M \times L} \]

Code rate

\[ R = \log_4 |C| \]

Decoder:

\[ \text{dec} : \mathbb{Z}^{N \times L} \rightarrow C \]

Error prob.

\[ P(\text{Err}) = P(\text{dec}(Y) \neq X) \]

Channel Capacity

Achievable rates

Code rate \( R \) is achievable, if there exists a code \( C \) of rate \( R \) with \( P(\text{Err}) \rightarrow 0 \), as \( ML \rightarrow \infty \)

Capacity: Supremum of achievable rates
Channel Model - Codes and Information Rates

Communication System:

Message

\[ W \rightarrow \text{Encoder} \]

\[ X \]

\[ \text{Draw & Distort} \]

\[ Y \]

\[ X_1 \rightarrow \text{GCTATGAGTACT} \]

\[ X_2 \rightarrow \text{ATGATTGACTCT} \]

\[ X_3 \rightarrow \text{GATGGCATAGCT} \]

\[ X_4 \rightarrow \text{ATGATTGACTCT} \]

\[ X_5 \rightarrow \text{GATGGCATACCT} \]

\[ Y_1 \rightarrow \text{ATAATTGAGTCT} \]

\[ Y_2 \rightarrow \text{GCTGGCATAGCT} \]

\[ Y_3 \rightarrow \text{GATAGCTTAGCT} \]

\[ Y_4 \rightarrow \text{ATGATTGACTCT} \]

\[ Y_5 \rightarrow \text{GATGGCATACCT} \]

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Model - Codes and Information Rates

Communication System:

Message

\[ W \rightarrow \text{Encoder} \rightarrow \text{Draw & Distort} \rightarrow \text{Decoder} \]

- **Encoder**: Takes a message and encodes it into a code.
- **Draw & Distort**: Introduces errors into the encoded data.
- **Decoder**: Decodes the distorted data back into a message.

**Code**: 
\[ C = \{ X(1), \ldots, X(4) \} \subset \mathbb{Z}^{M \times L} \]

**Code rate**: 
\[ R = \log_4 |C|_{\text{ML}} \]

**Error prob.**: 
\[ P(\text{Err}) = P(\text{dec}(Y) \neq X) \]

**Channel Capacity**: The supremum of achievable rates.

- **Achievable rates**: The code rate \( R \) is achievable if there exists a code \( C \) of rate \( R \) with \( P(\text{Err}) \rightarrow 0 \) as \( \text{ML} \rightarrow \infty \).

**Capacity**: 
\[ C \text{ is the capacity of the channel.} \]

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Model - Codes and Information Rates

Communication System:

Message $W$ → Encoder → $X$ Draw & Distort $Y$ → Decoder → $\widehat{W}$

- **Message** $W$
- **Encoder**
- **Decoder**

**Code:**
- $C = \{X_1, \ldots, X_{ML}\} \subset \mathbb{Z}^{M \times L}$
- **Code rate** $R = \log_4|C|_{ML}$

**Decoder:**
- $\text{dec} : \mathbb{Z}^{N \times L} \rightarrow C$
- **Error prob.** $P(\text{Err}) = P(\text{dec}(Y) \neq X)$

**Channel Capacity:**
- Achievable rates $R$ is achievable, if there exists a code $C$ of rate $R$ with $P(\text{Err}) \rightarrow 0$, as $ML \rightarrow \infty$
- **Capacity**: Supremum of achievable rates
Channel Model - Codes and Information Rates

Communication System:

Message $W$ \rightarrow Encoder \rightarrow Decoder \rightarrow \hat{W}

Code:

- $C = \{X(1), \ldots, X(4^{MLR})\} \subset \mathbb{Z}_4^{M \times L}$
- Code rate $R = \frac{\log_4 |C|}{ML}$
Channel Model - Codes and Information Rates

Communication System:

Message $W$ \(\xrightarrow{\text{Encoder}}\) \(\text{Decoder} \xrightarrow{\hat{W}}\)

Code:
- $\mathcal{C} = \{X(1), \ldots, X(4^{MLR})\} \subset \mathbb{Z}_4^{M \times L}$
- Code rate $R = \frac{\log_4 |\mathcal{C}|}{ML}$

Decoder:
- $\text{dec} : \mathbb{Z}_4^{N \times L} \rightarrow \mathcal{C}$
- Error prob. $P(\text{Err}) = P(\text{dec}(Y) \neq X)$

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Model - Codes and Information Rates

Communication System:

Message

\( W \) → Encoder

\[ X \]

Draw & Distort

\[ Y \]

\( X_1 \)

\( X_2 \)

\( X_3 \)

\( X_4 \)

\( Y_1 \)

\( Y_2 \)

\( Y_3 \)

\( Y_4 \)

\( Y_5 \)

\( \hat{W} \) → Decoder

Code:

- \( C = \{X(1), \ldots, X(4^{MLR})\} \subseteq \mathbb{Z}_4^{M \times L} \)
- Code rate \( R = \frac{\log_4 |C|}{ML} \)

Decoder:

- dec : \( \mathbb{Z}_4^{N \times L} \rightarrow C \)
- Error prob. \( P(\text{Err}) = P(\text{dec}(Y) \neq X) \)

Channel Capacity

Achievable rates

Code rate \( R \) is achievable, if there exists a code \( C \) of rate \( R \) with \( P(\text{Err}) \rightarrow 0 \), as \( ML \rightarrow \infty \)
Channel Model - Codes and Information Rates

Communication System:

Message

\[ W \rightarrow \text{Encoder} \rightarrow \text{Decoder} \rightarrow \widehat{W} \]

Message

\[ W \]

**Code:**

- \( \mathcal{C} = \{ X(1), \ldots, X(4^{MLR}) \} \subset \mathbb{Z}_4^{M \times L} \)
- Code rate \( R = \frac{\log_4 |\mathcal{C}|}{ML} \)

**Decoder:**

- \( \text{dec} : \mathbb{Z}_4^{N \times L} \rightarrow \mathcal{C} \)
- Error prob. \( P(\text{Err}) = P(\text{dec}(Y) \neq X) \)

Channel Capacity

**Achievable rates**

Code rate \( R \) is achievable, if there exists a code \( \mathcal{C} \) of rate \( R \) with \( P(\text{Err}) \rightarrow 0 \), as \( ML \rightarrow \infty \)

- **Capacity**: Supremum of achievable rates

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Related Work

[Mitzenmacher, ”On the Theory and Practice of Data Recovery with Multiple Versions,” 2006]

- Capacity of binomial/multi-draw channel
Related Work

[Mitzenmacher, ”On the Theory and Practice of Data Recovery with Multiple Versions,” 2006]

- Capacity of binomial/multi-draw channel

[Heckel et al., ”Fundamental Limits of DNA Storage Systems,” 2017]

- Introduced channel model with no errors $p = 0$
Related Work


- Capacity of binomial/multi-draw channel

[Heckel et al., “Fundamental Limits of DNA Storage Systems,” 2017]

- Introduced channel model with no errors $p = 0$
- Computed capacity $C = (1 - e^{-c})(1 - \beta)$
Related Work

[Mitzenmacher, ”On the Theory and Practice of Data Recovery with Multiple Versions,” 2006]

- Capacity of binomial/multi-draw channel

[Heckel et al., ”Fundamental Limits of DNA Storage Systems,” 2017]

- Introduced channel model with no errors $p = 0$
- Computed capacity $C = (1 - e^{-c})(1 - \beta)$

[Shomorony et al., ”Capacity of the Noisy Shuffling Channel,” 2019]

- Similar channel - each sequence is drawn exactly once with errors
Related Work

[Mitzenmacher, "On the Theory and Practice of Data Recovery with Multiple Versions," 2006]

- Capacity of binomial/multi-draw channel

[Heckel et al., "Fundamental Limits of DNA Storage Systems," 2017]

- Introduced channel model with no errors $p = 0$
- Computed capacity $C = (1 - e^{-c})(1 - \beta)$

[Shomorony et al., "Capacity of the Noisy Shuffling Channel," 2019]

- Similar channel - each sequence is drawn exactly once with errors
- Computed capacity $C = 1 - H(p) - \beta$
Related Work

[Mitzenmacher, "On the Theory and Practice of Data Recovery with Multiple Versions," 2006]

• Capacity of binomial/multi-draw channel

[Heckel et al., "Fundamental Limits of DNA Storage Systems," 2017]

• Introduced channel model with no errors $p = 0$
• Computed capacity $C = (1 - e^{-c})(1 - \beta)$

[Shomorony et al., "Capacity of the Noisy Shuffling Channel," 2019]

• Similar channel - each sequence is drawn exactly once with errors
• Computed capacity $C = 1 - H(p) - \beta$

[Shomorony et al., "DNA-based storage: Models and fundamental limits", 2021]

• Generalization to Bernoulli drawing distributions with success prob. $q$
Related Work

[Mitzenmacher, ”On the Theory and Practice of Data Recovery with Multiple Versions,” 2006]

- Capacity of binomial/multi-draw channel

[Heckel et al., ”Fundamental Limits of DNA Storage Systems,” 2017]

- Introduced channel model with no errors $p = 0$
- Computed capacity $C = (1 - e^{-c})(1 - \beta)$

[Shomorony et al., ”Capacity of the Noisy Shuffling Channel,” 2019]

- Similar channel - each sequence is drawn exactly once with errors
- Computed capacity $C = 1 - H(p) - \beta$

[Shomorony et al., ”DNA-based storage: Models and fundamental limits”, 2021]

- Generalization to Bernoulli drawing distributions with success prob. $q$
- $C = (1 - q)(1 - H(p) - \beta)$
Related Work

[Lenz et al., ”An Upper Bound on the Capacity of the DNA Storage Channel,” 2019]

- Upper bound on capacity
Related Work

[Lenz et al., ”An Upper Bound on the Capacity of the DNA Storage Channel,” 2019]
  • Upper bound on capacity

[Lenz et al., ”Achieving the Capacity of the DNA Storage Channel,” 2020]
  • Prove achievability of the capacity in [Lenz et al., 2019]
Related Work

[Lenz et al., ”An Upper Bound on the Capacity of the DNA Storage Channel,” 2019]

• Upper bound on capacity

[Lenz et al., ”Achieving the Capacity of the DNA Storage Channel,” 2020]

• Prove achievability of the capacity in [Lenz et al., 2019]

[Weinberger, Merhav, ”The DNA Storage Channel: Capacity and Error Probability Bounds,” 2021]

• Generalization to asymmetric channels
Related Work

[Lenz et al., ”An Upper Bound on the Capacity of the DNA Storage Channel,” 2019]
- Upper bound on capacity

[Lenz et al., ”Achieving the Capacity of the DNA Storage Channel,” 2020]
- Prove achievability of the capacity in [Lenz et al., 2019]

[Weinberger, Merhav, ”The DNA Storage Channel: Capacity and Error Probability Bounds,” 2021]
- Generalization to asymmetric channels
- Computation of error probabilities
Preliminaries - Channel Model Revisited

Alternative Channel Formulation

\[ X \]

\[ X_1 \text{ GCTATGAGTACT} \]

\[ X_2 \text{ ATGATTGACTCT} \]

\[ X_3 \text{ GATGGCATAGCT} \]

Challenges

- Draws of the multi-draw channels are random
- Permutation of the sequences

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Preliminaries - Channel Model Revisited

Alternative Channel Formulation

\[ X \quad \text{Multi-draw} \]

\[ X_1 \quad \text{GCTATGAGTACT} \]
\[ D_1 = 0 \]

\[ X_2 \quad \text{ATGATTGACTCT} \]
\[ D_2 = 2 \]
\[ \text{ATAATTGAGTCT} \]
\[ \text{ATGATTGACTCT} \]

\[ X_3 \quad \text{GATGGCATAGCT} \]
\[ D_3 = 3 \]
\[ \text{GCTGGCATAAGCT} \]
\[ \text{GATAGCTTAGCT} \]
\[ \text{GATGGCATAACCT} \]

Challenges

- Draws of the multi-draw channels are random
- Permutation of the sequences

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Preliminaries - Channel Model Revisited

Alternative Channel Formulation

<table>
<thead>
<tr>
<th></th>
<th>$X$</th>
<th>Multi-draw</th>
<th>Permute</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>GCTATGAGTACT</td>
<td>$D_1 = 0$</td>
<td>Y_1: ATAATTGAGTCT</td>
</tr>
<tr>
<td>$X_2$</td>
<td>ATGATTGACTCT</td>
<td>$D_2 = 2$</td>
<td>Y_2: GCTGGCATAGCT</td>
</tr>
<tr>
<td>$X_3$</td>
<td>GATGGCATAGCT</td>
<td>$D_3 = 3$</td>
<td>Y_3: GATAGCTTAGCT</td>
</tr>
</tbody>
</table>

Challenges

- Draws of the multi-draw channels are random
- Permutation of the sequences

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Preliminaries - Channel Model Revisited

Alternative Channel Formulation

\[
\begin{align*}
X & \quad \text{Multi-draw} & \quad \text{Permute} \\
X_1 & \quad \text{GCTATGAGTACT} & \quad D_1 = 0 & \quad \text{Y} \\
X_2 & \quad \text{ATGATTGACTCT} & \quad D_2 = 2 & \quad \text{Y}_1 \quad \text{ATATTGAGTCT} \\
X_3 & \quad \text{GATGGCATAGCT} & \quad D_3 = 3 & \quad \text{Y}_2 \quad \text{GCTGGCATACCT} \\
& & \quad \text{ATGATTGACTCT} & \quad \text{Y}_3 \quad \text{GATAGCTTAGCT} \\
& & \quad \text{GCTGGCATAGCT} & \quad \text{Y}_4 \quad \text{ATGATTGACTCT} \\
& & \quad \text{GATAGCTTAGCT} & \quad \text{Y}_5 \quad \text{GATGGCATACCT} \\
& & \quad \text{GATGGCATACCT} & \\
\end{align*}
\]

Challenges

- Draws of the multi-draw channels are random

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Preliminaries - Channel Model Revisited

Alternative Channel Formulation

\[ X \]

\[ X_1 \quad \text{GCTATGAGTACT} \]

\[ D_1 = 0 \]

\[ X_2 \quad \text{ATGATTGACTCT} \]

\[ D_2 = 2 \]

\[ X_3 \quad \text{GATGGCATAGCT} \]

\[ D_3 = 3 \]

\[ Y \]

\[ Y_1 \quad \text{ATAATTGAGTCT} \]

\[ Y_2 \quad \text{GCTGGCATAGCT} \]

\[ Y_3 \quad \text{GATAGCTTAGCT} \]

\[ Y_4 \quad \text{ATGATTGACTCT} \]

\[ Y_5 \quad \text{GATGGCATAACCT} \]

Permute

Challenges

- Draws of the multi-draw channels are random
- Permutation of the sequences

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Preliminaries - Multi-draw Channel [Mitzenmacher, 2006]

Let $X = \text{ATGATTGACTCT}$ be the input sequence. It is mapped through a multi-draw channel to $Y_1, Y_2, \ldots, Y_d$.

- $Y_1 = \text{ATAATTGAGTCT}$
- $Y_2 = \text{ATGATTGACTCT}$
- $Y_d = \text{ATAATTAACTTT}$

This illustrates the transmission of DNA sequences through a channel with substitution errors.
Preliminaries - Multi-draw Channel [Mitzenmacher, 2006]

Input: $X$

Output: $d$ output sequences $Y_1, \ldots, Y_d$

$q$-ary symmetric channels: $Y_i = X + E_i$
Preliminaries - Multi-draw Channel [Mitzenmacher, 2006]

- Input: \( X \)
- Output: \( d \) output sequences \( Y_1, \ldots, Y_d \)
- \( q \)-ary symmetric channels: \( Y_i = X + E_i \)
Preliminaries - Multi-draw Channel [Mitzenmacher, 2006]

- Capacity (d draws, error probability p)

\[ C(d, p) \]

\[ \begin{align*}
  &\text{\texttt{ATAATTGACTCT}} \\
  &\text{\texttt{ATAATTGACTCT}} \\
  &\vdots \\
  &\text{\texttt{ATAATTAACTTT}} \\
\end{align*} \]

\[ Y_d \]

\[ X \]

\[ \text{\texttt{ATGATTGACTCT}} \]

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Capacity

Draw Distribution

- Recall:
  - $D_i$: Number of draws of sequence $i$
Channel Capacity

Draw Distribution

• Recall:
  ▶ $D_i$: Number of draws of sequence $i$
  ▶ $c = N/M$: Sequencing depth (average number of draws per sequence)
• $D_i \rightarrow \text{Poi}(c)$ (Poissonization)
Channel Capacity

Draw Distribution

• Recall:
  ▶ $D_i$: Number of draws of sequence $i$
  ▶ $c = N/M$: Sequencing depth (average number of draws per sequence)
• $D_i \rightarrow \text{Poi}(c)$ (Poissonization)

Channel Capacity

Theorem: Channel Capacity

Given $2\beta < 1 - H_4(2p)$, the capacity is

$$C(c, \beta, p) = \sum_{d=0}^{\infty} \text{Poi}(c,d)C_{\text{Mul}}(d, p) - \beta(1 - e^{-c})$$

Lenz,Siegel,Wachter-Zeh,Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Capacity - Parameter Range

Parameter Range

\[ 2\beta < 1 - H_4(2p) \]

- Entropy function \( H_4(p) = -(1 - p) \log_4(1 - p) - p \log_4 \left( \frac{p}{3} \right) \)
Channel Capacity - Parameter Range

Parameter Range

\[ 2\beta < 1 - H_4(2p) \]

- Entropy function \( H_4(p) = -(1 - p) \log_4(1 - p) - p \log_4 \left( \frac{p}{3} \right) \)

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Capacity - Discussion

Capacity

\[ C'(c, \beta, p) = \sum_{d=0}^{\infty} Poi(c, d) C_{\text{Mul}}(d, p) - \beta (1 - e^{-c}) \]
Capacity

\[C(c, \beta, p) = \sum_{d=0}^{\infty} \text{Poi}(c, d) C_{\text{Mul}}(d, p) - \beta(1 - e^{-c})\]

\[\beta = 1/20\]
Channel Capacity - Discussion

Capacity

\[ C(c, \beta, p) = \sum_{d=0}^{\infty} \text{Poi}(c,d)C_{\text{Mul}}(d, p) - \beta(1 - e^{-c}) \]
Channel Capacity - Discussion

Capacity

\[ C'(c, \beta, p) = \sum_{d=0}^{\infty} \text{Poi}(c, d) C_{\text{Mul}}(d, p) - \beta (1 - e^{-c}) \]

- No errors \((p = 0)\) [Heckel 2017]
Channel Capacity - Discussion

Capacity

\[ C(c, \beta, p) = \sum_{d=0}^{\infty} \text{Poi}(c, d) C_{\text{Mul}}(d, p) - \beta (1 - e^{-c}) \]

- No errors \( (p = 0) \) [Heckel 2017]

\[ C(c, \beta, p) = (1 - e^{-c})(1 - \beta) \]
Channel Capacity - Discussion

Capacity

\[
C(c, \beta, p) = \sum_{d=0}^{\infty} \text{Poi}(c,d) C_{\text{Mul}}(d,p) - \beta(1 - e^{-c})
\]

- No errors \((p = 0)\) [Heckel 2017]

\[
C(c, \beta, p) = (1 - e^{-c})(1 - \beta)
\]

- Many draws \((c \to \infty)\)
Capacity

\[ C'(c, \beta, p) = \sum_{d=0}^{\infty} \text{Poi}(c, d) C_{\text{Mul}}(d, p) - \beta (1 - e^{-c}) \]

- No errors \((p = 0)\) [Heckel 2017]

\[ C(c, \beta, p) = (1 - e^{-c})(1 - \beta) \]

- Many draws \((c \to \infty)\)

\[ C(c, \beta, p) = 1 - \beta \]
Channel Capacity - Discussion

Capacity

\[ C(c, \beta, p) = \sum_{d=0}^{\infty} \text{Poi}(c, d) C_{\text{Mul}}(d, p) - \beta (1 - e^{-c}) \]

- No errors \((p = 0)\) [Heckel 2017]

\[ C(c, \beta, p) = (1 - e^{-c})(1 - \beta) \]

- Many draws \((c \to \infty)\)

\[ C(c, \beta, p) = 1 - \beta \]

- Long sequences \((\beta \to 0)\)
Channel Capacity - Discussion

Capacity

\[ C(c, \beta, p) = \sum_{d=0}^{\infty} \text{Poi}(c, d) C_{\text{Mul}}(d, p) - \beta (1 - e^{-c}) \]

- No errors \((p = 0)\) [Heckel 2017]

\[ C(c, \beta, p) = (1 - e^{-c})(1 - \beta) \]

- Many draws \((c \to \infty)\)

\[ C(c, \beta, p) = 1 - \beta \]

- Long sequences \((\beta \to 0)\)

\[ C(c, \beta, p) = \sum_{d=0}^{\infty} \text{Poi}(c, d) C_{\text{Mul}}(d, p) \]
Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

- Storage rate: $R_s = \log_2 |C|/ML$
Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

- Storage rate: $R_s = \log_2 |C|/ML$
- Recovery rate: $R_r = \log_2 |C|/NL$
Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

- Storage rate: \( R_s = \log_2 |C|/ML \)
- Recovery rate: \( R_r = \log_2 |C|/NL \)
Channel Capacity - Storage and Recovery Rate Tradeoff

**Storage and Recovery Rate Tradeoff**

- Storage rate: \( R_s = \frac{\log_2 |C|}{ML} \)
- Recovery rate: \( R_r = \frac{\log_2 |C|}{NL} \)
Channel Capacity - Storage and Recovery Rate Tradeoff

**Storage and Recovery Rate Tradeoff**

- **Storage rate:** \( R_s = \log_2 |C|/ML \)
- **Recovery rate:** \( R_r = \log_2 |C|/NL \)
Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

- Storage rate: \( R_s = \log_2 |C| / ML \)
- Recovery rate: \( R_r = \log_2 |C| / NL \)

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

- Storage rate: \( R_s = \frac{\log_2 |C|}{ML} \)
- Recovery rate: \( R_r = \frac{\log_2 |C|}{NL} \)

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

- Storage rate: \( R_s = \frac{\log_2 |C|}{ML} \)
- Recovery rate: \( R_r = \frac{\log_2 |C|}{NL} \)

Lenz, Siegel, Wachter-Zeh, Yaakobi "On the Capacity of DNA-based Data Storage under Substitution Errors"
Summary & Outlook

Summary

- Capacity of the DNA storage channel under substitution errors

Outlook

- Challenges for efficiently encodable/decodable schemes
  - Each input sequence goes through channel that is unknown a priori
  - How to combat the indexing problem?
  - Modified concatenated codes [Lenz et al., 2020]
- Insertions/deletions
- Runlength constraints, balanced GC content
Summary & Outlook

Summary

- Capacity of the DNA storage channel under substitution errors
- Connection with multi-draw channels

Outlook

- Challenges for efficiently encodable/decodable schemes
  - Each input sequence goes through channel that is unknown a priori
  - How to combat the indexing problem?
  - Modified concatenated codes [Lenz et al., 2020]

- Insertions/deletions
- Runlength constraints, balanced GC content
Summary & Outlook

Summary

• Capacity of the DNA storage channel under substitution errors
• Connection with multi-draw channels
• Storage/Recovery rate tradeoff

Outlook

• Challenges for efficiently encodable/decodable schemes
  ▶ Each input sequence goes through channel that is unknown a priori
  ▶ How to combat the indexing problem?
  ▶ Modified concatenated codes [Lenz et al., 2020]

• Insertions/deletions
• Runlength constraints, balanced GC content
Summary & Outlook

Summary

• Capacity of the DNA storage channel under substitution errors
• Connection with multi-draw channels
• Storage/Recovery rate tradeoff

Outlook

• Challenges for efficiently encodable/decodable schemes
Summary & Outlook

Summary

- Capacity of the DNA storage channel under substitution errors
- Connection with multi-draw channels
- Storage/Recovery rate tradeoff

Outlook

- Challenges for efficiently encodable/decodable schemes
  - Each input sequence goes through channel that is unknown a priori
Summary & Outlook

Summary

• Capacity of the DNA storage channel under substitution errors
• Connection with multi-draw channels
• Storage/Recovery rate tradeoff

Outlook

• Challenges for efficiently encodable/decodable schemes
  ▶ Each input sequence goes through channel that is unknown a priori
  ▶ How to combat the indexing problem?
Summary & Outlook

Summary

- Capacity of the DNA storage channel under substitution errors
- Connection with multi-draw channels
- Storage/Recovery rate tradeoff

Outlook

- Challenges for efficiently encodable/decodable schemes
  - Each input sequence goes through channel that is unknown a priori
  - How to combat the indexing problem?
  - Modified concatenated codes [Lenz et al., 2020]
Summary & Outlook

Summary

- Capacity of the DNA storage channel under substitution errors
- Connection with multi-draw channels
- Storage/Recovery rate tradeoff

Outlook

- Challenges for efficiently encodable/decodable schemes
  - Each input sequence goes through channel that is unknown a priori
  - How to combat the indexing problem?
  - Modified concatenated codes [Lenz et al., 2020]
- Insertions/deletions
Summary & Outlook

Summary

• Capacity of the DNA storage channel under substitution errors
• Connection with multi-draw channels
• Storage/Recovery rate tradeoff

Outlook

• Challenges for efficiently encodable/decodable schemes
  ▶ Each input sequence goes through channel that is unknown a priori
  ▶ How to combat the indexing problem?
  ▶ Modified concatenated codes [Lenz et al., 2020]
• Insertions/deletions
• Runlength constraints, balanced GC content
Summary & Outlook

Summary

- Capacity of the DNA storage channel under substitution errors
- Connection with multi-draw channels
- Storage/Recovery rate tradeoff

Outlook

- Challenges for efficiently encodable/decodable schemes
  - Each input sequence goes through channel that is unknown a priori
  - How to combat the indexing problem?
  - Modified concatenated codes [Lenz et al., 2020]
- Insertions/deletions
- Runlength constraints, balanced GC content

Thank you!