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

Andreas Lenz[1], Paul H. Siegel[2], Antonia Wachter-Zeh[1], Eitan Yaakobi[3]

[1]Institute for Communications Engineering
Technical University of Munich

[2]Department of Electrical and Computer Engineering
University of California

[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: $10^9\,\mathrm{GB/mm^3}$
- Tape: $10 - 100\,\mathrm{GB/mm^3}$

Data Storage in DNA

Never obsolete

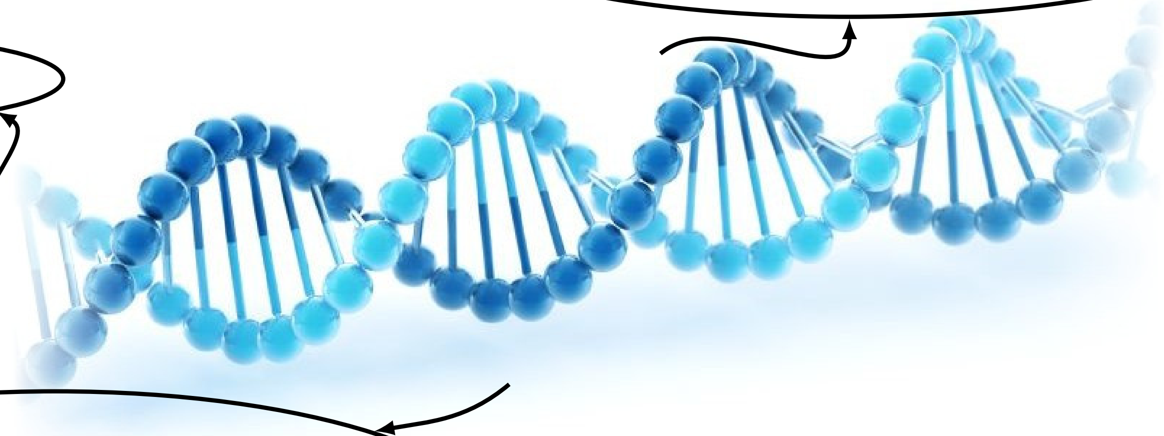
High density data storage

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

Data Storage in DNA

Long term data storage
(DNA from mammoths)

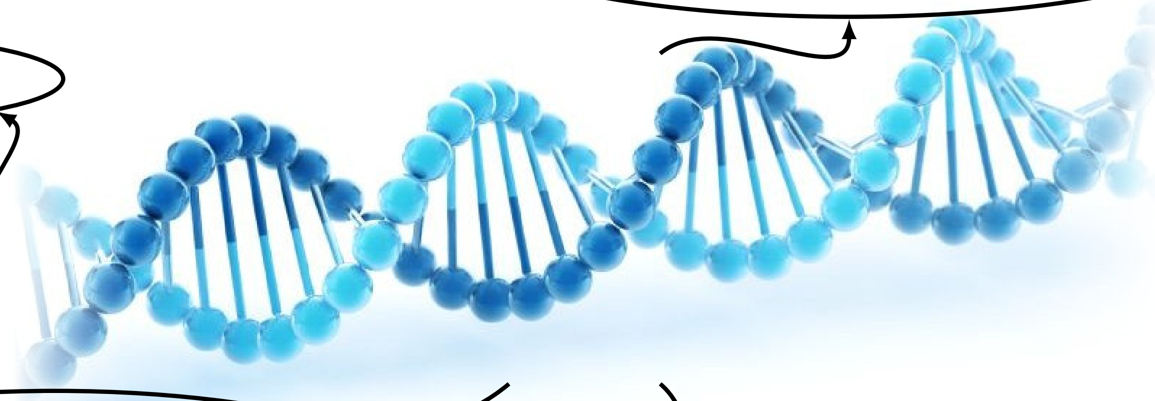
Never obsolete



High density data storage

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

Data Storage in DNA



Long term data storage
(DNA from mammoths)

Never obsolete

High density data storage

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

Easily duplicatable (PCR)

Data Storage in DNA

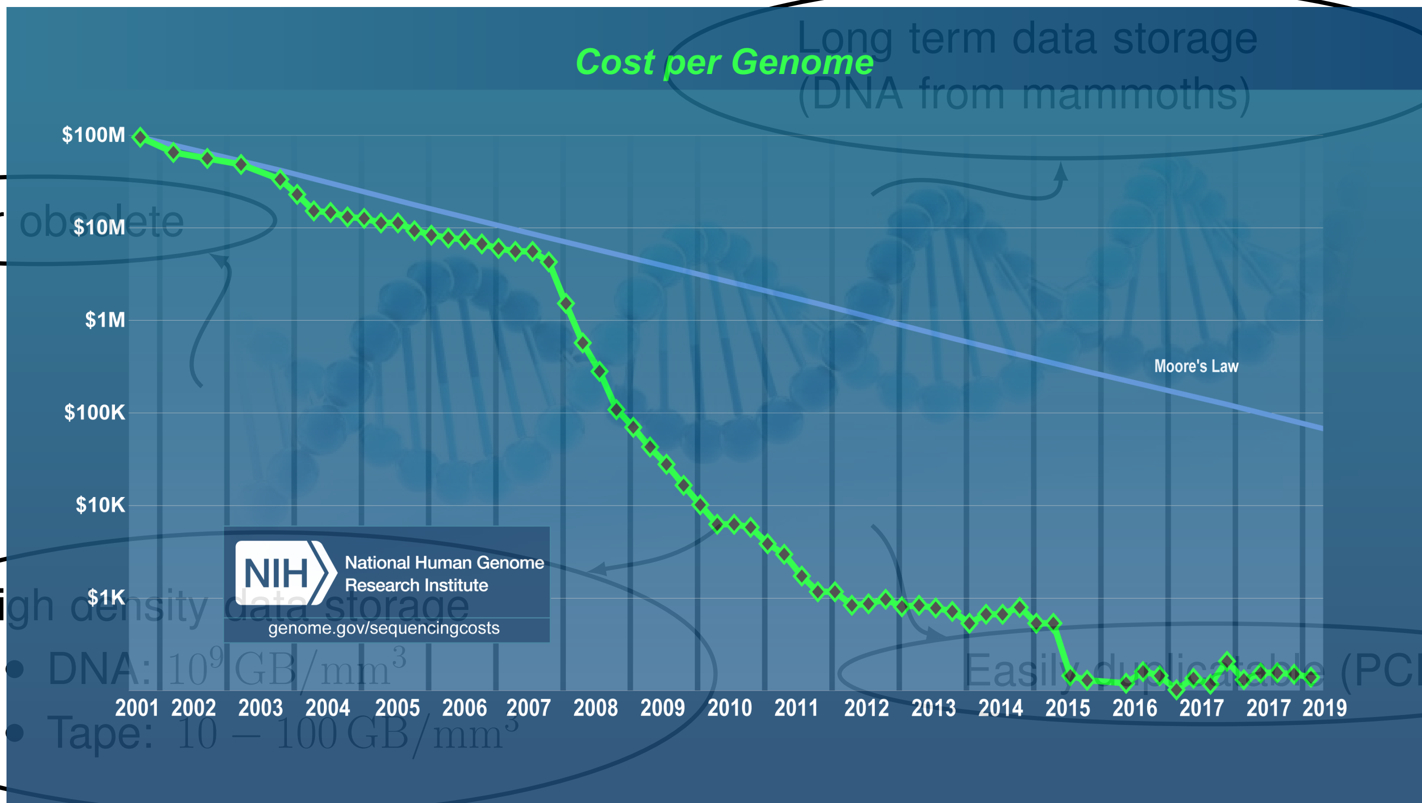
Long term data storage (DNA from mammoths)

Never obsolete

High density data storage

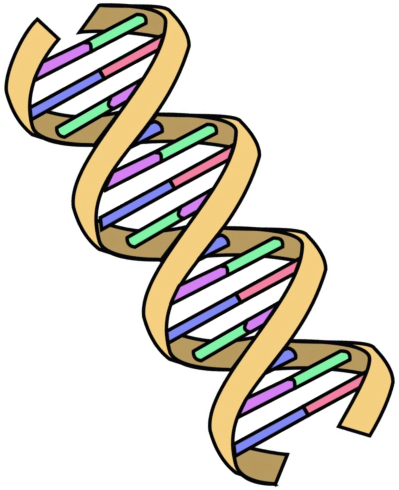
- DNA: $10^9\,\mathrm{GB/mm^3}$
- Tape: $10 - 100\,\mathrm{GB/mm^3}$

Easily duplicatable (PCR)

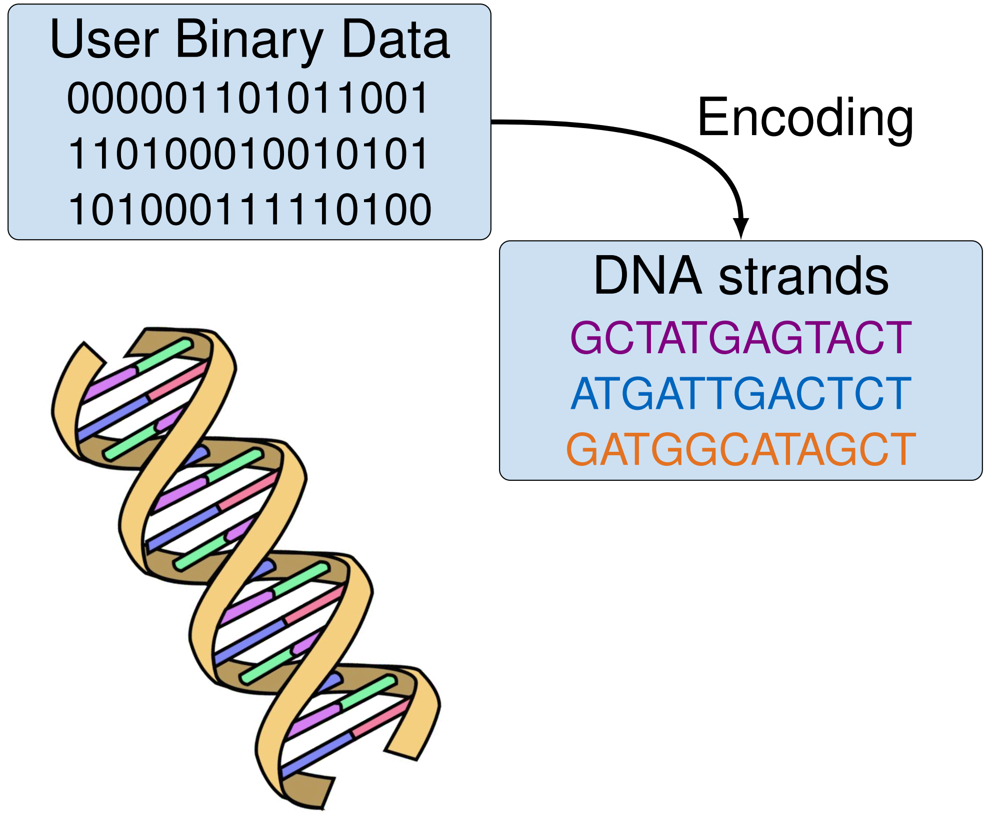


Data Storage in DNA

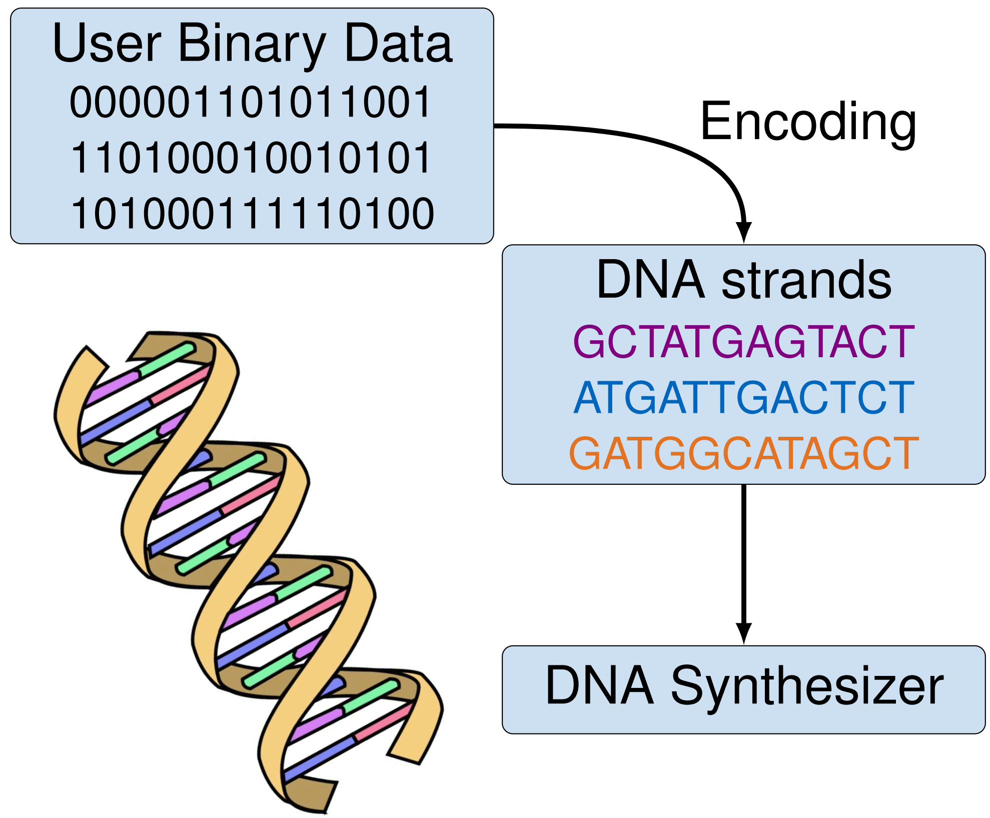
User Binary Data
000001101011001
110100010010101
101000111110100



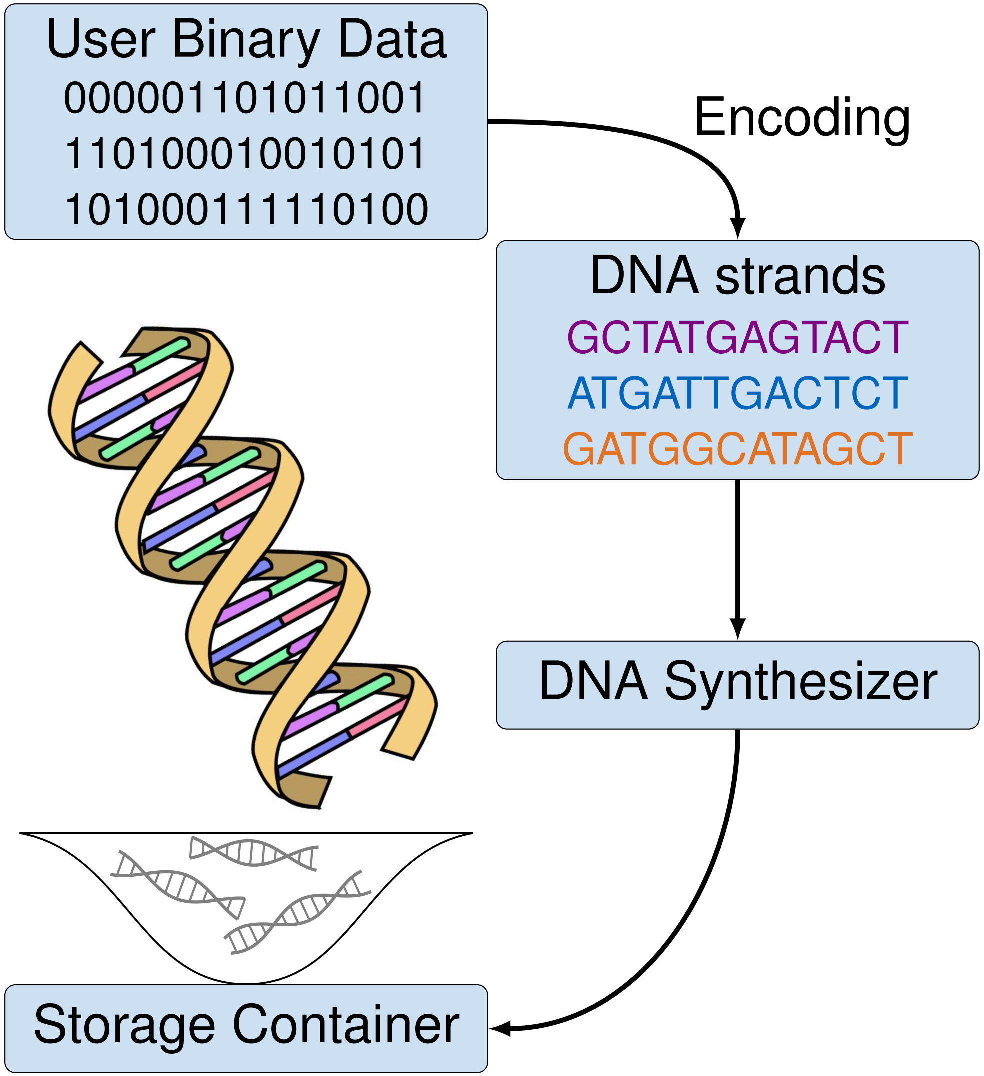
Data Storage in DNA



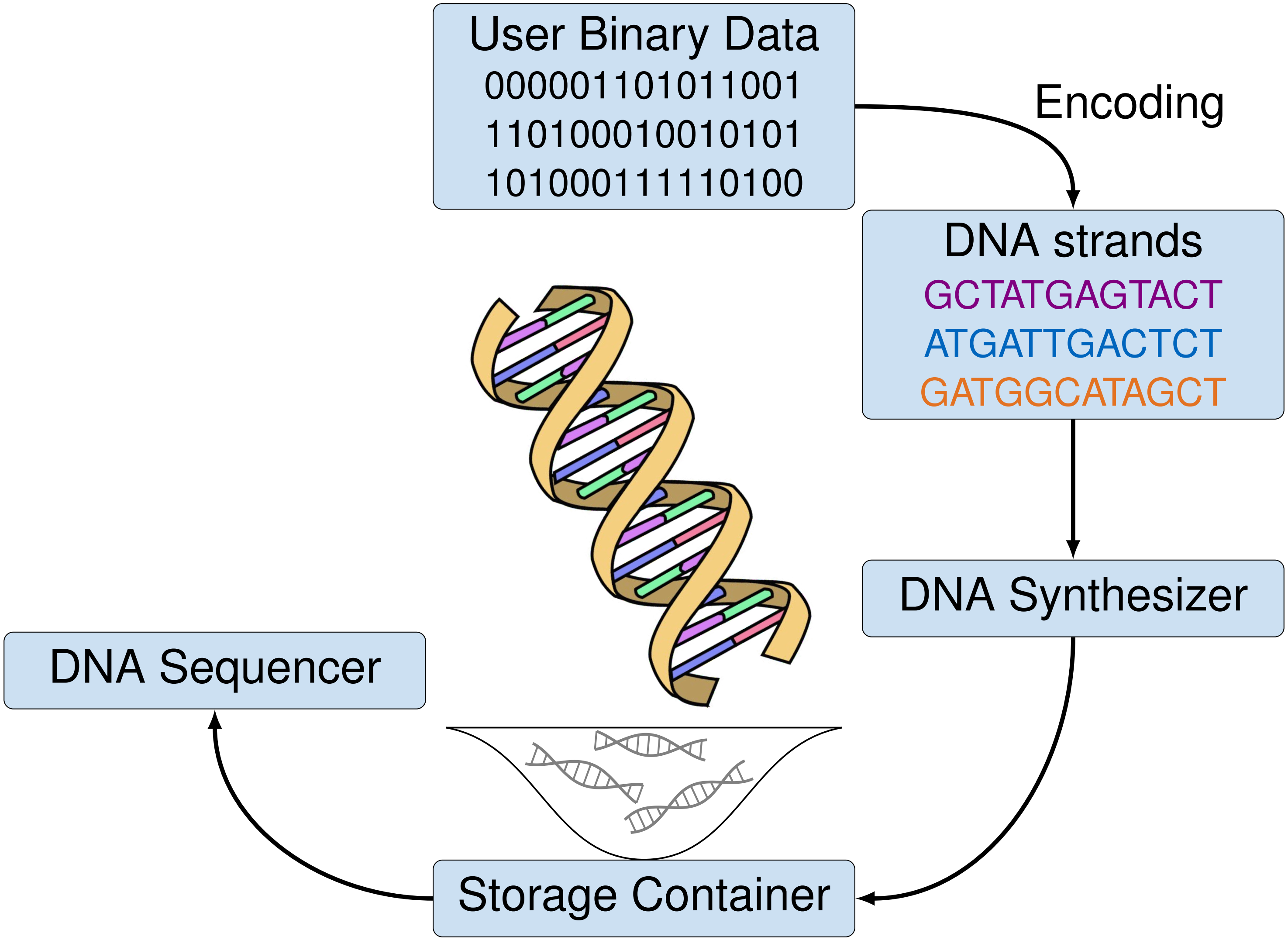
Data Storage in DNA



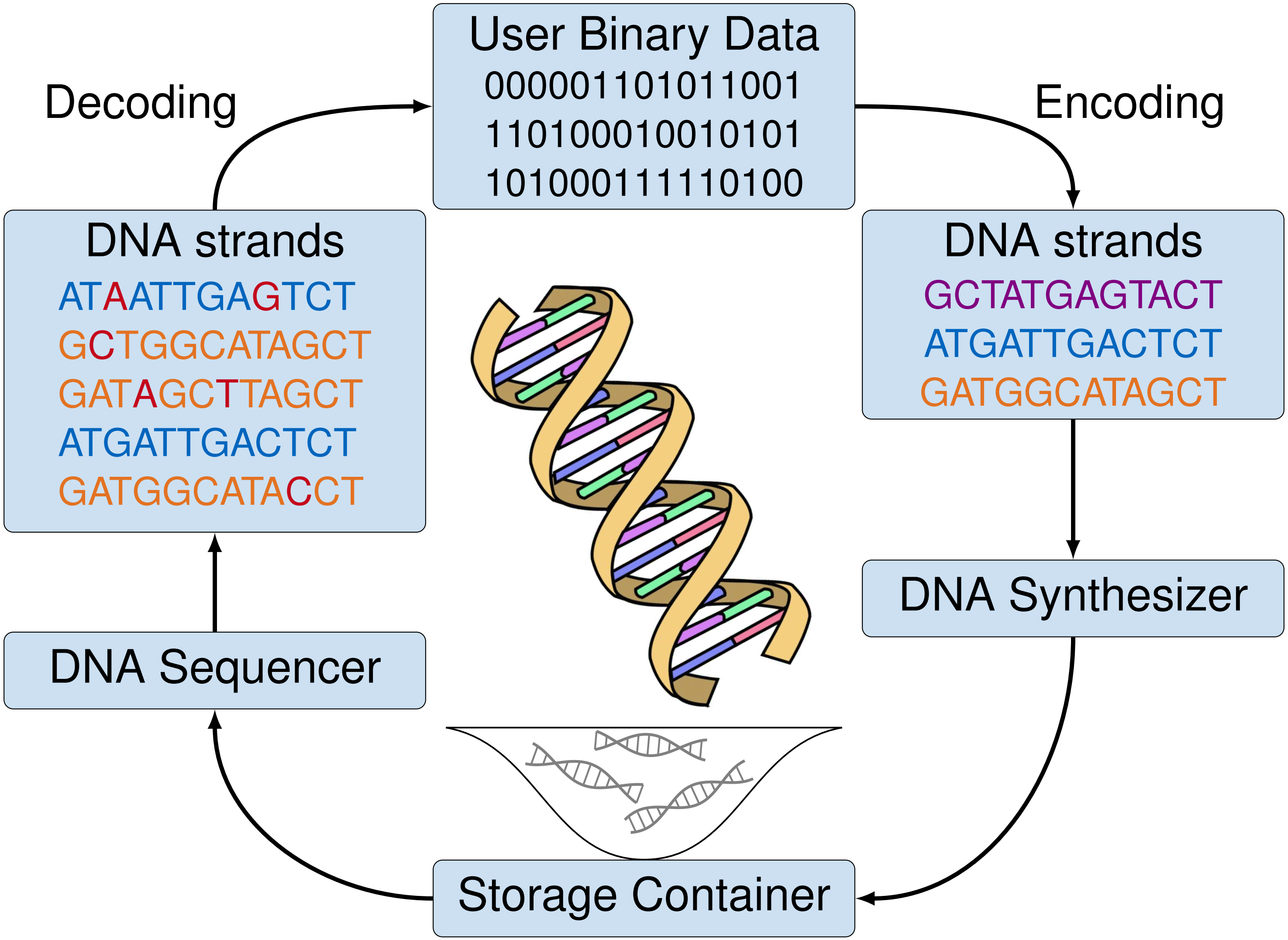
Data Storage in DNA



Data Storage in DNA



Data Storage in DNA



Channel Model

Y

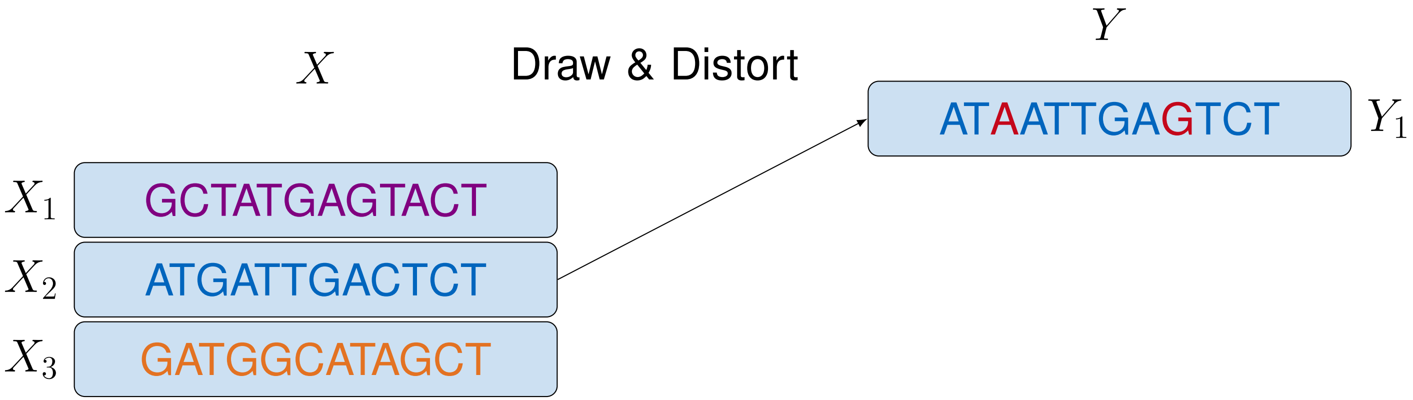
X Draw & Distort

X_1 GCTATGAGTACT

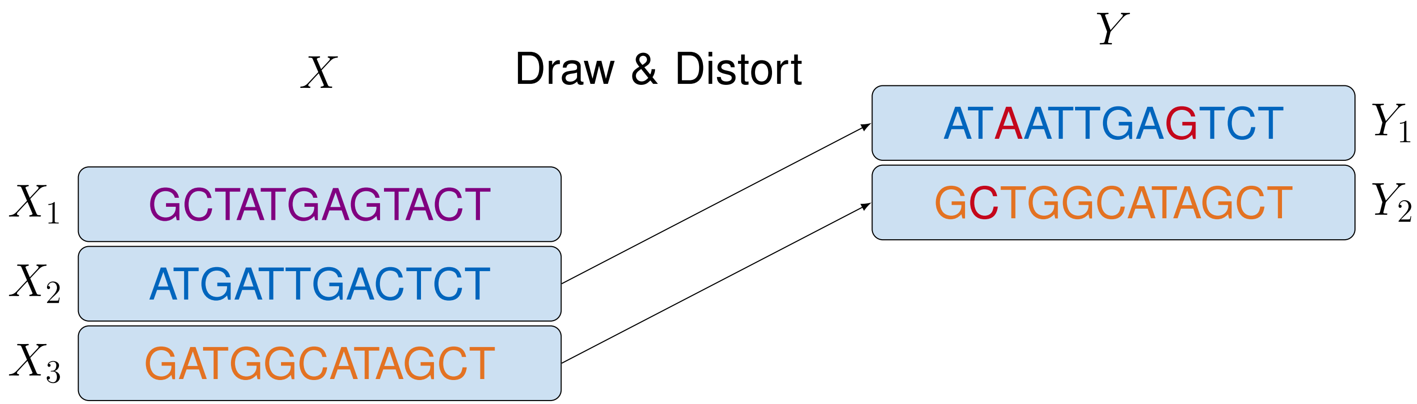
X_2 ATGATTGACTCT

X_3 GATGGCATAGCT

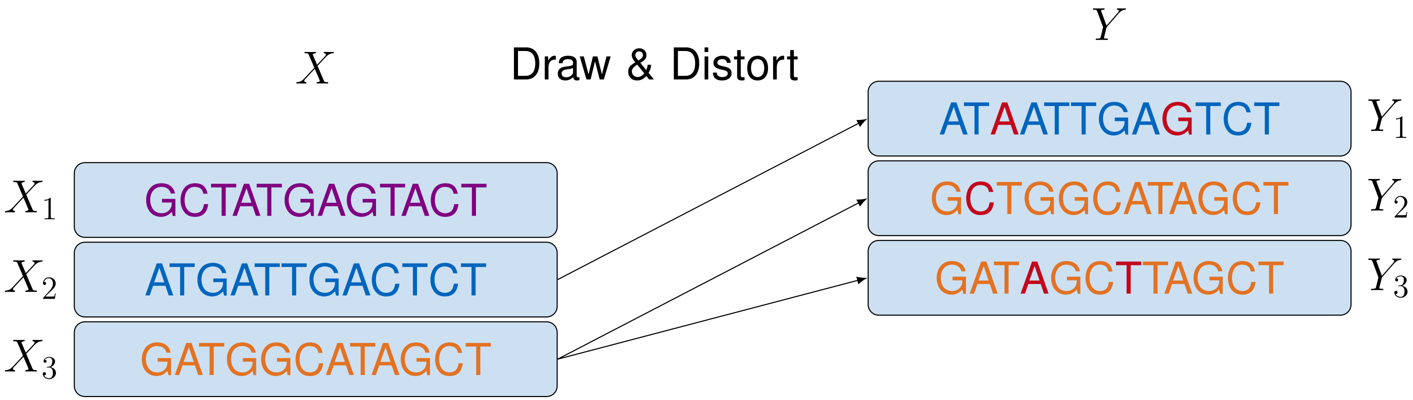
Channel Model



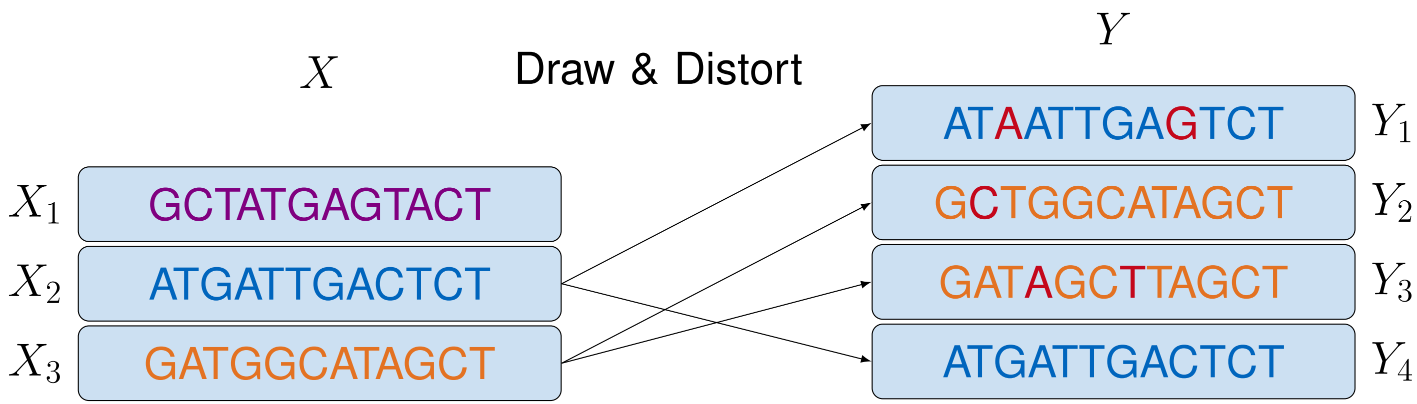
Channel Model



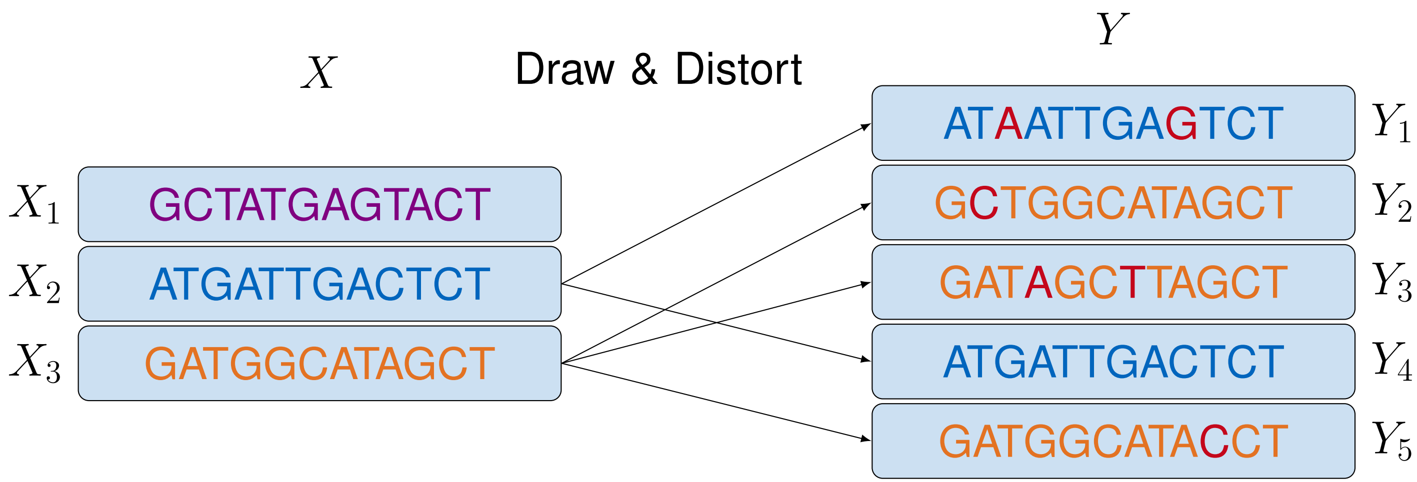
Channel Model



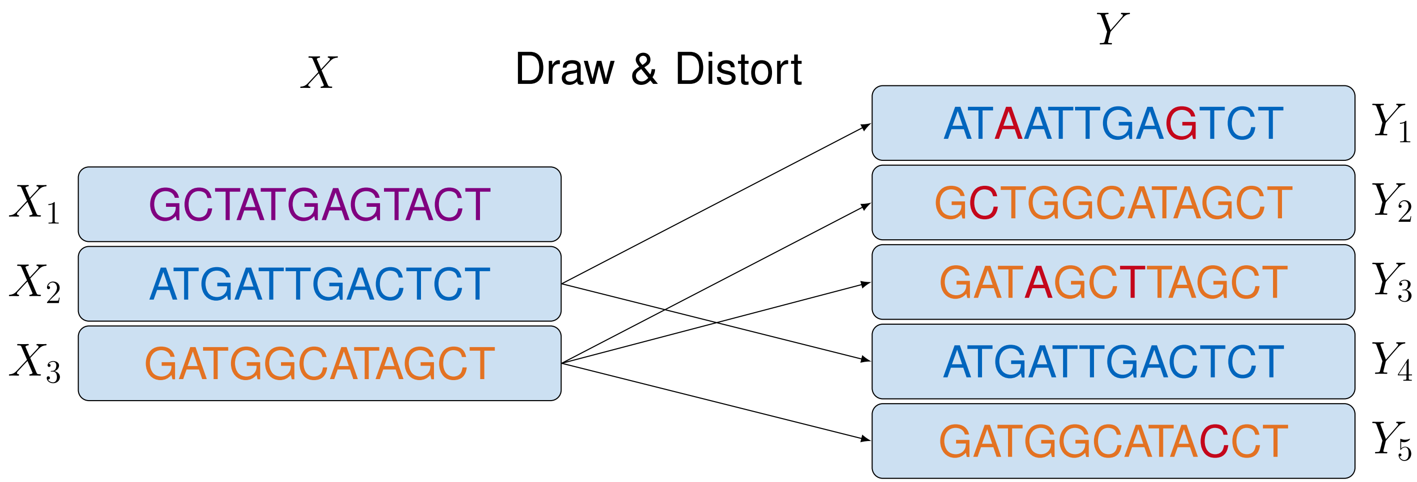
Channel Model



Channel Model



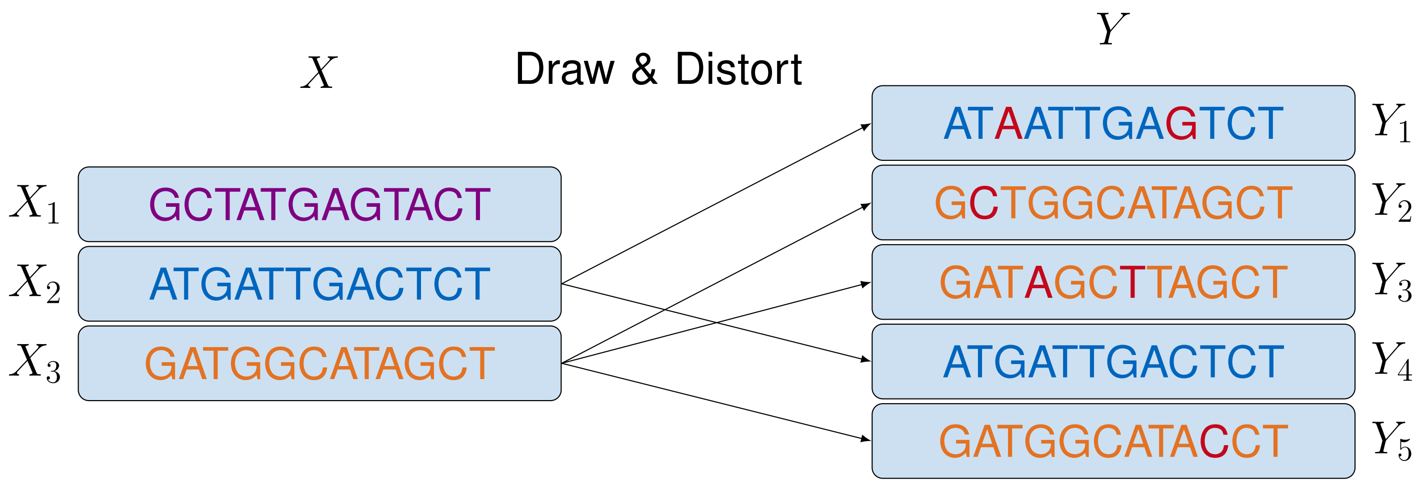
Channel Model



$$Y_j = X_{I_j} + E_j$$

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

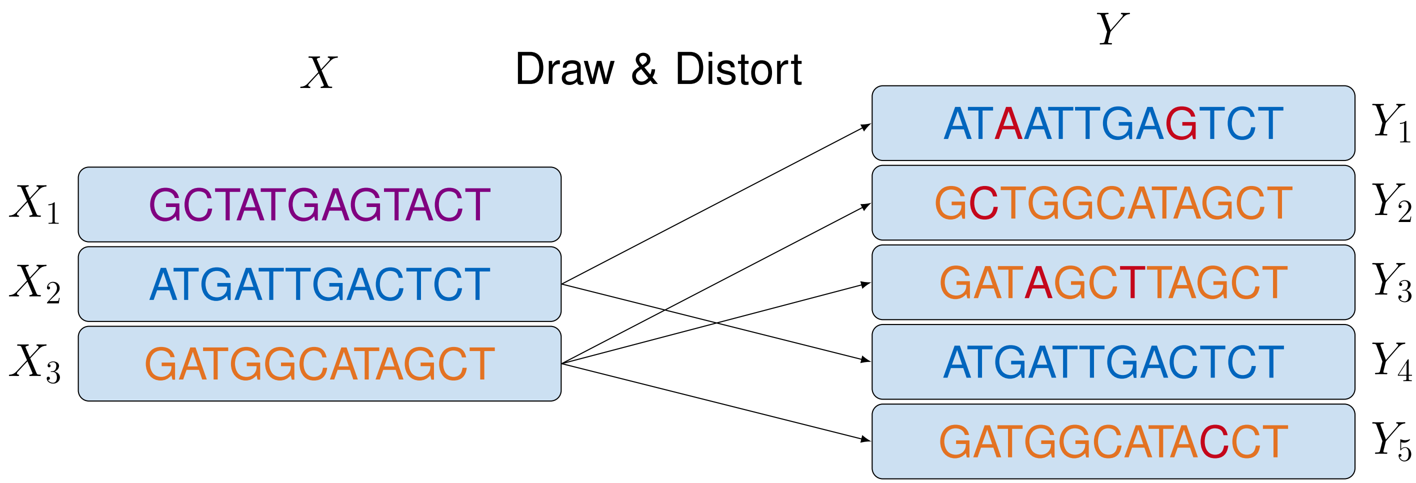
Channel Model



$$Y_j = X_{I_j} + E_j$$

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

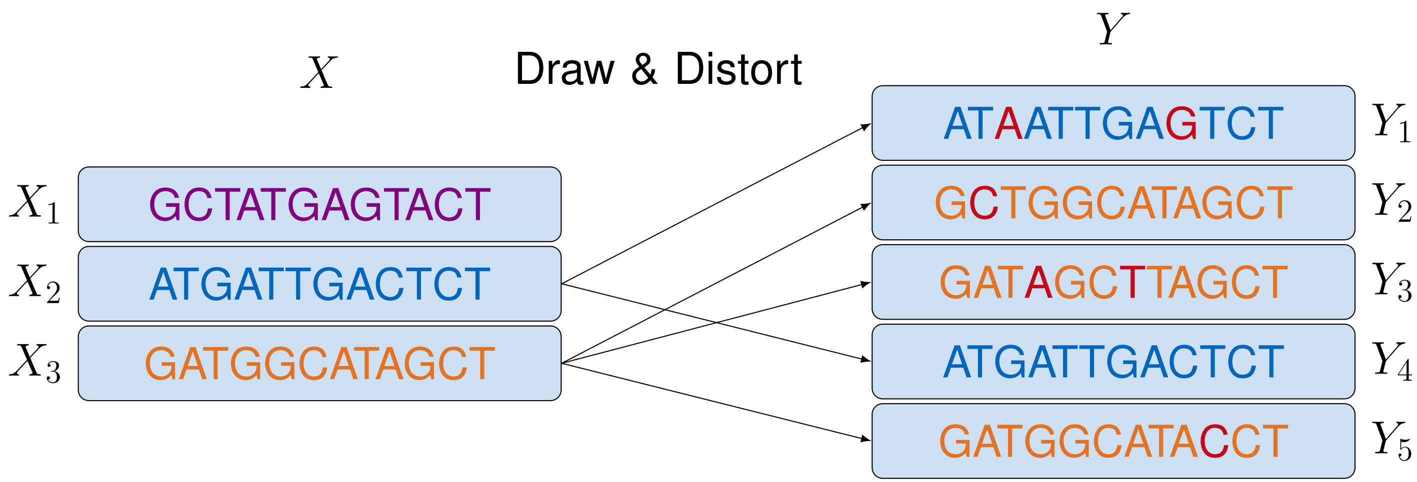
Channel Model



Channel Input:

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

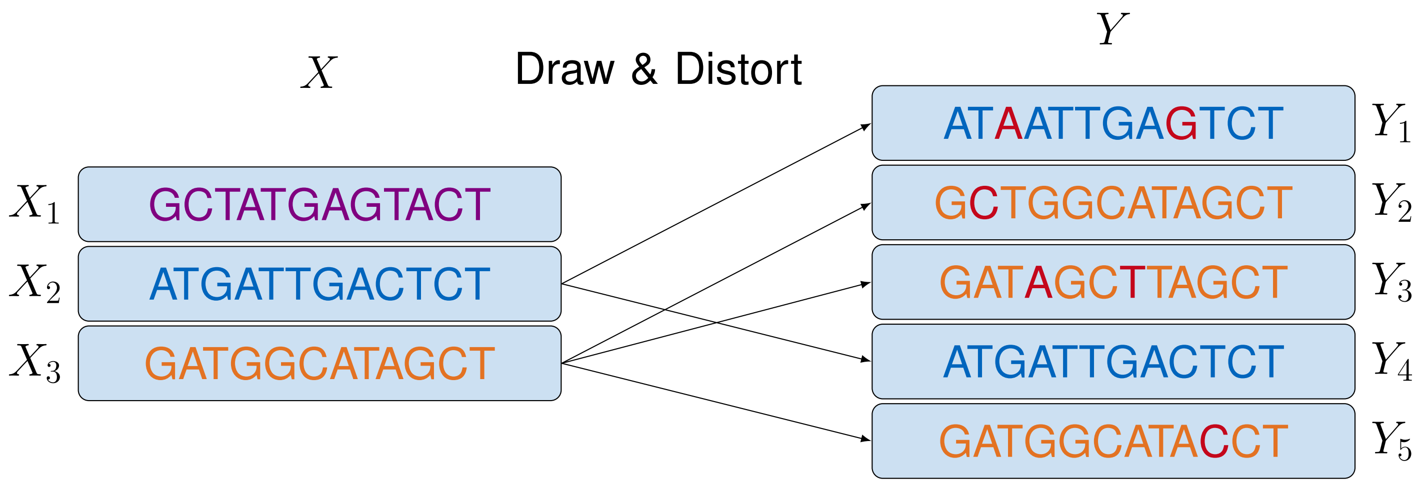
Channel Model



Channel Input:

- M Sequences, each of length L
- $X = (X_1, \ldots, X_M)$
- $\beta = \log_4 M / L$

Channel Model



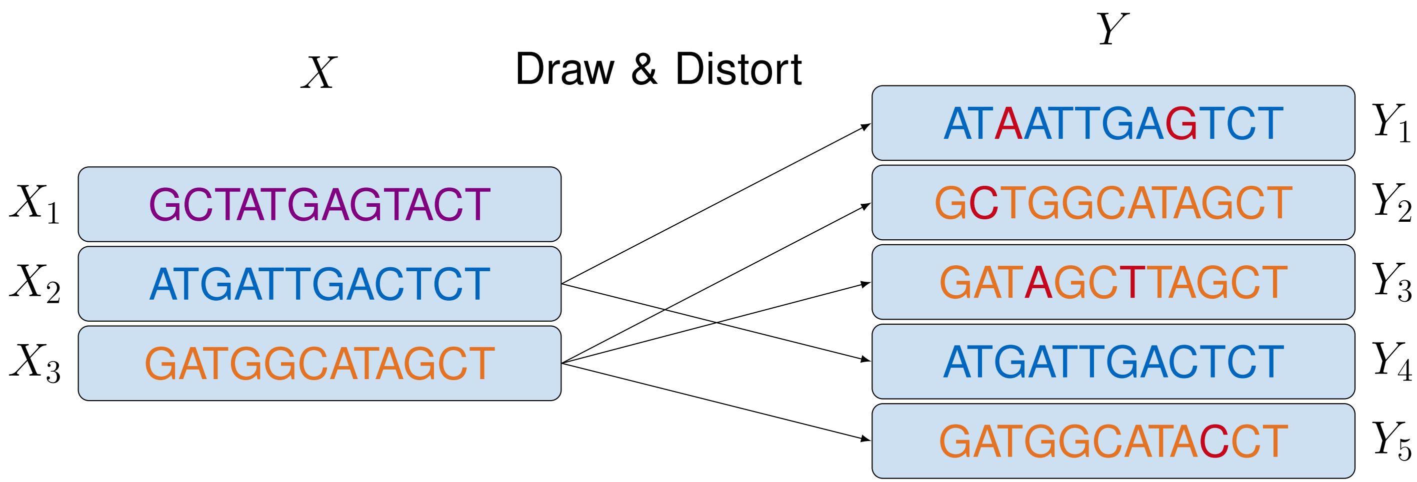
Channel Input:

- M Sequences, each of length L
- $X = (X_1, \ldots, X_M)$
- $\beta = \log_4 M / L$

Channel Output:

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

Channel Model



Channel Input:

- M Sequences, each of length L
- $X = (X_1, \ldots, X_M)$
- $\beta = \log_4 M / L$

Channel Output:

- N sequences, each of length L
- $Y = (Y_1, \ldots, Y_N)$
- $c = N / M$

Channel Model - Codes and Information Rates

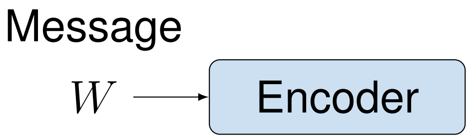
Communication System:

Message

W

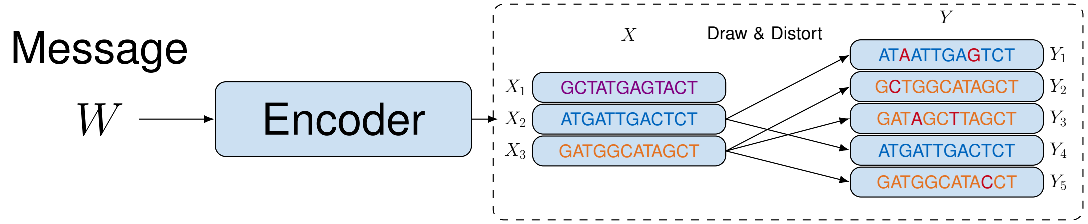
Channel Model - Codes and Information Rates

Communication System:



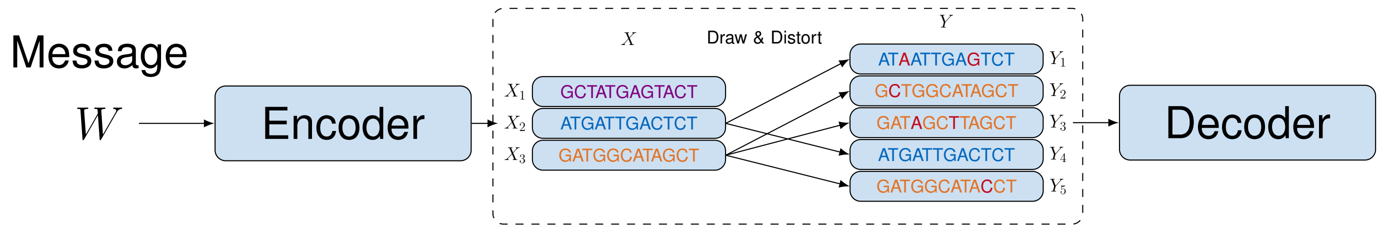
Channel Model - Codes and Information Rates

Communication System:



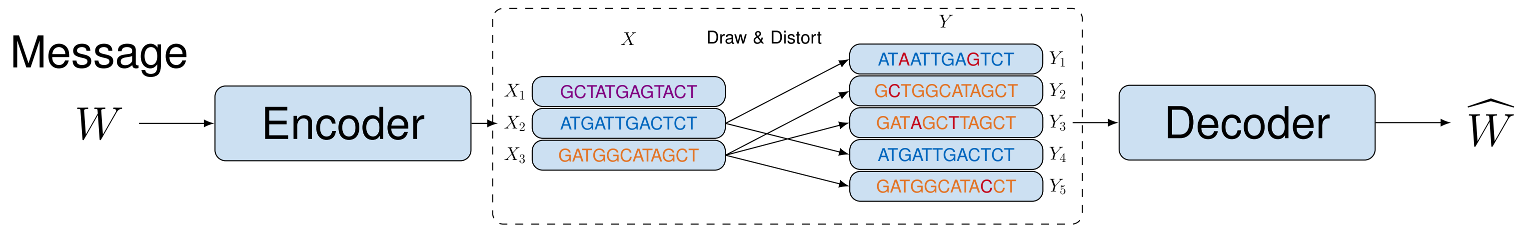
Channel Model - Codes and Information Rates

Communication System:



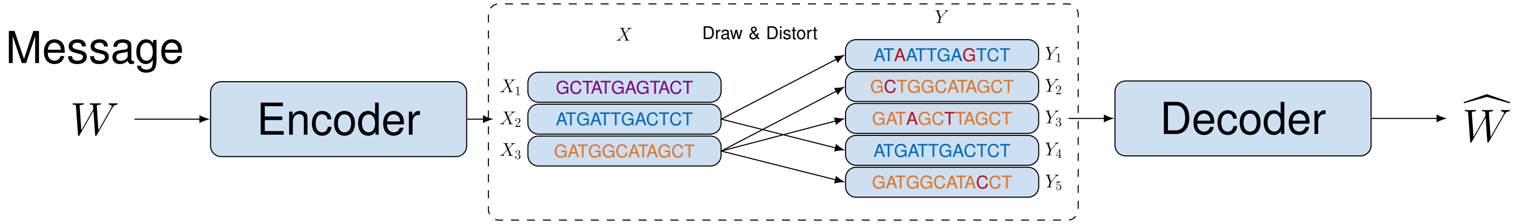
Channel Model - Codes and Information Rates

Communication System:



Channel Model - Codes and Information Rates

Communication System:

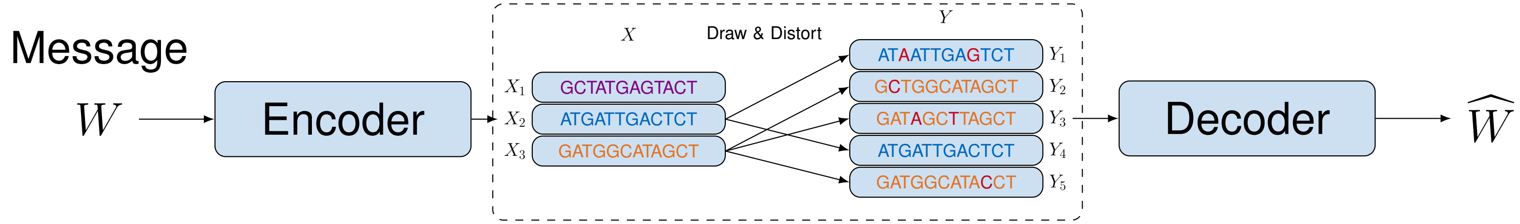


Code:

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

Channel Model - Codes and Information Rates

Communication System:



Code:

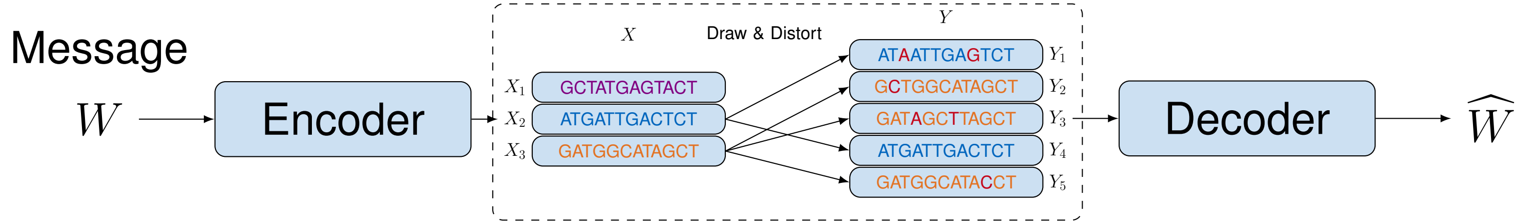
- $\mathcal{C} = \{X(1), \ldots, X(4^{MLR})\} \subset \mathbb{Z}_4^{M \times L}$
- Code rate $R = \frac{\log_4 |\mathcal{C}|}{ML}$

Decoder:

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

Channel Model - Codes and Information Rates

Communication System:



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:

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

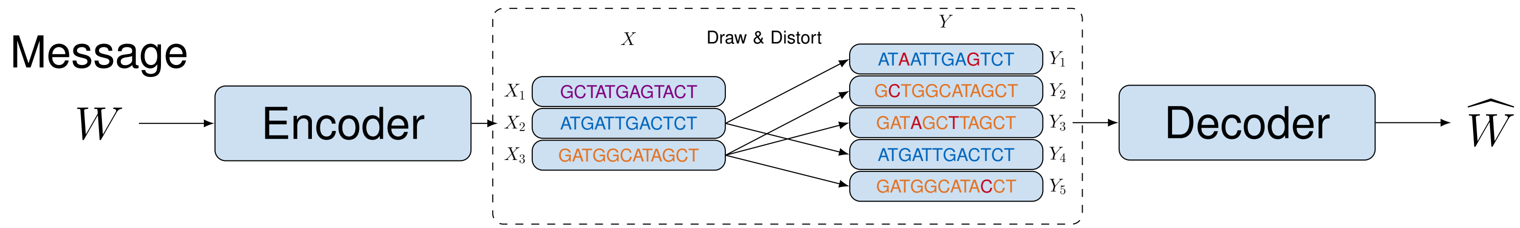
Channel Capacity

Achievable rates

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

Channel Model - Codes and Information Rates

Communication System:



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:

- dec : $\mathbb{Z}_4^{N \times L} \mapsto \mathcal{C}$
- 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 $P(\text{Err}) \to 0$, as $ML \to \infty$

- **Capacity**: Supremum of achievable rates

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

[**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 $C = (1 - \mathrm{e}^{-c})(1 - \beta)$

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 $C = (1 - \mathrm{e}^{-c})(1 - \beta)$

[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 $C = (1 - \mathrm{e}^{-c})(1 - \beta)$

[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 $C = (1 - \mathrm{e}^{-c})(1 - \beta)$

[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 $C = (1 - \mathrm{e}^{-c})(1 - \beta)$

[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

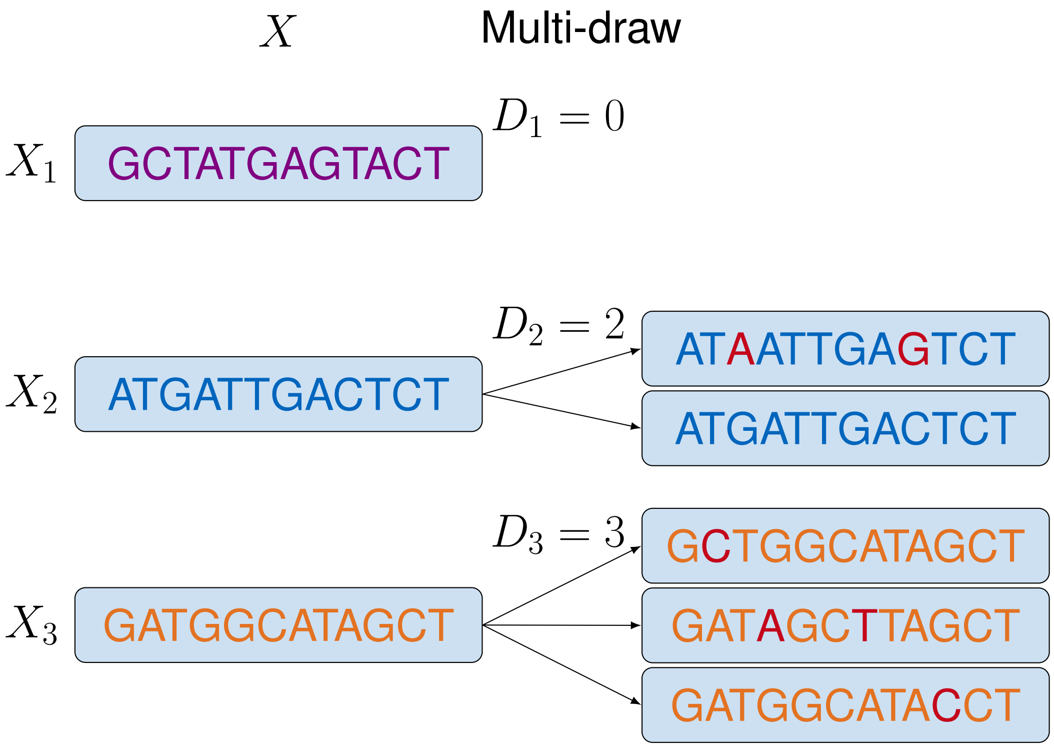
X_1 GCTATGAGTACT

X_2 ATGATTGACTCT

X_3 GATGGCATAGCT

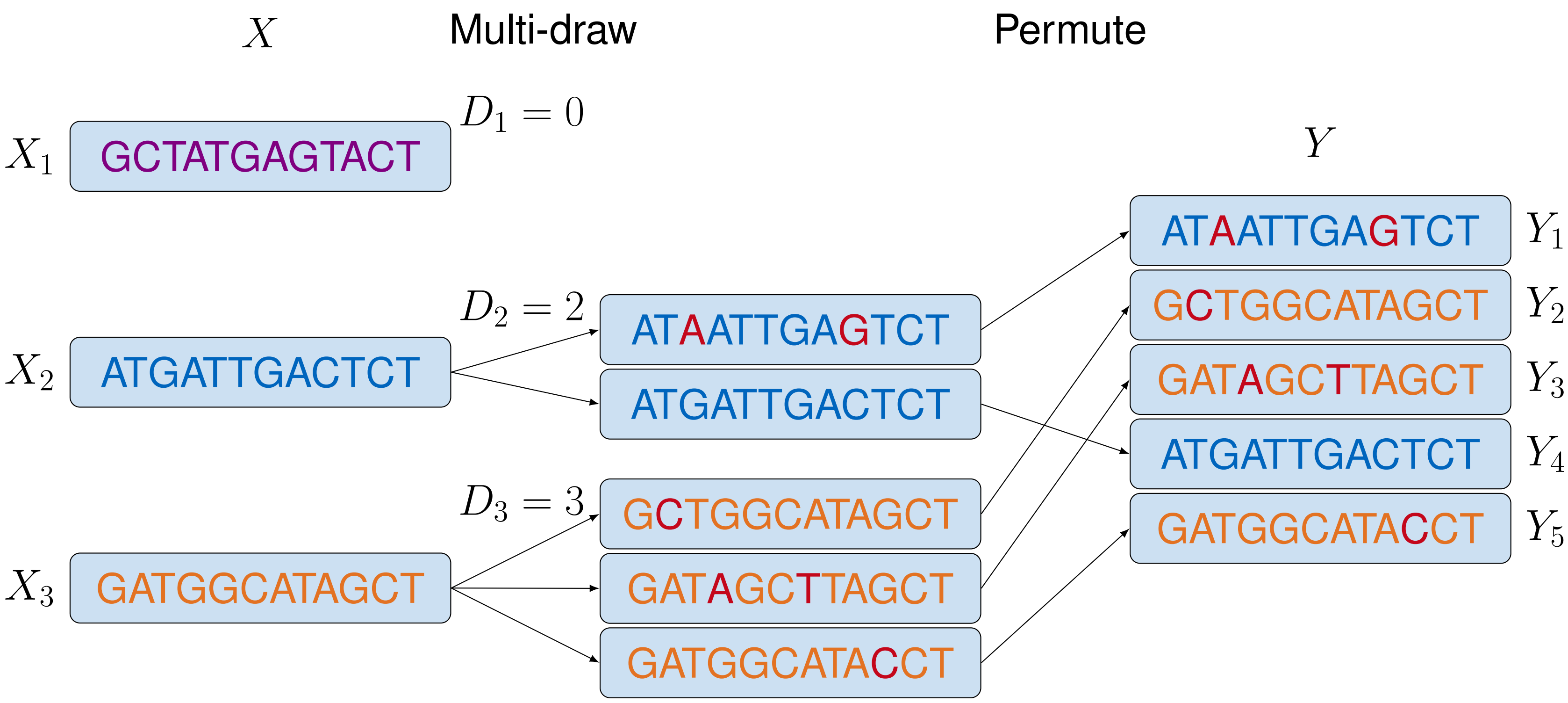
Preliminaries - Channel Model Revisited

Alternative Channel Formulation



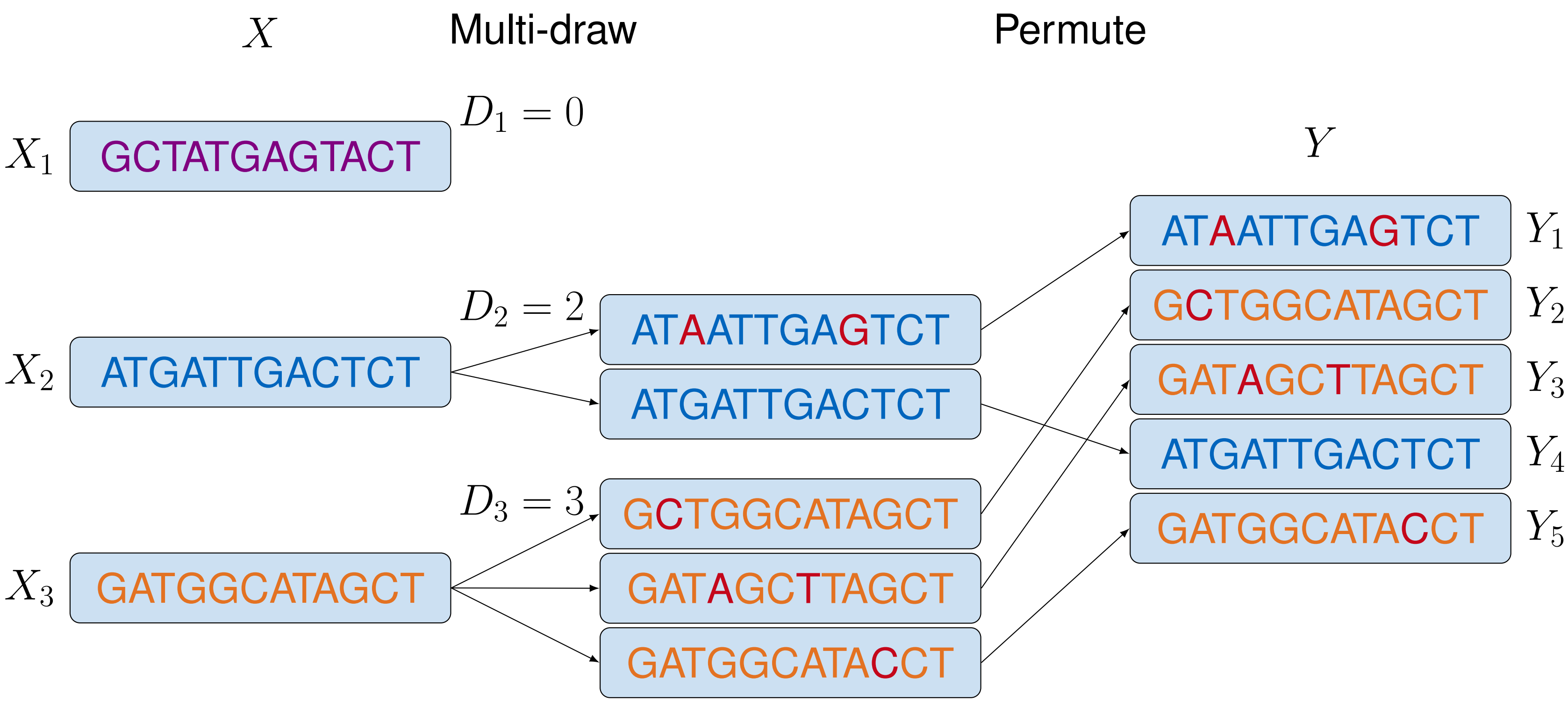
Preliminaries - Channel Model Revisited

Alternative Channel Formulation



Preliminaries - Channel Model Revisited

Alternative Channel Formulation

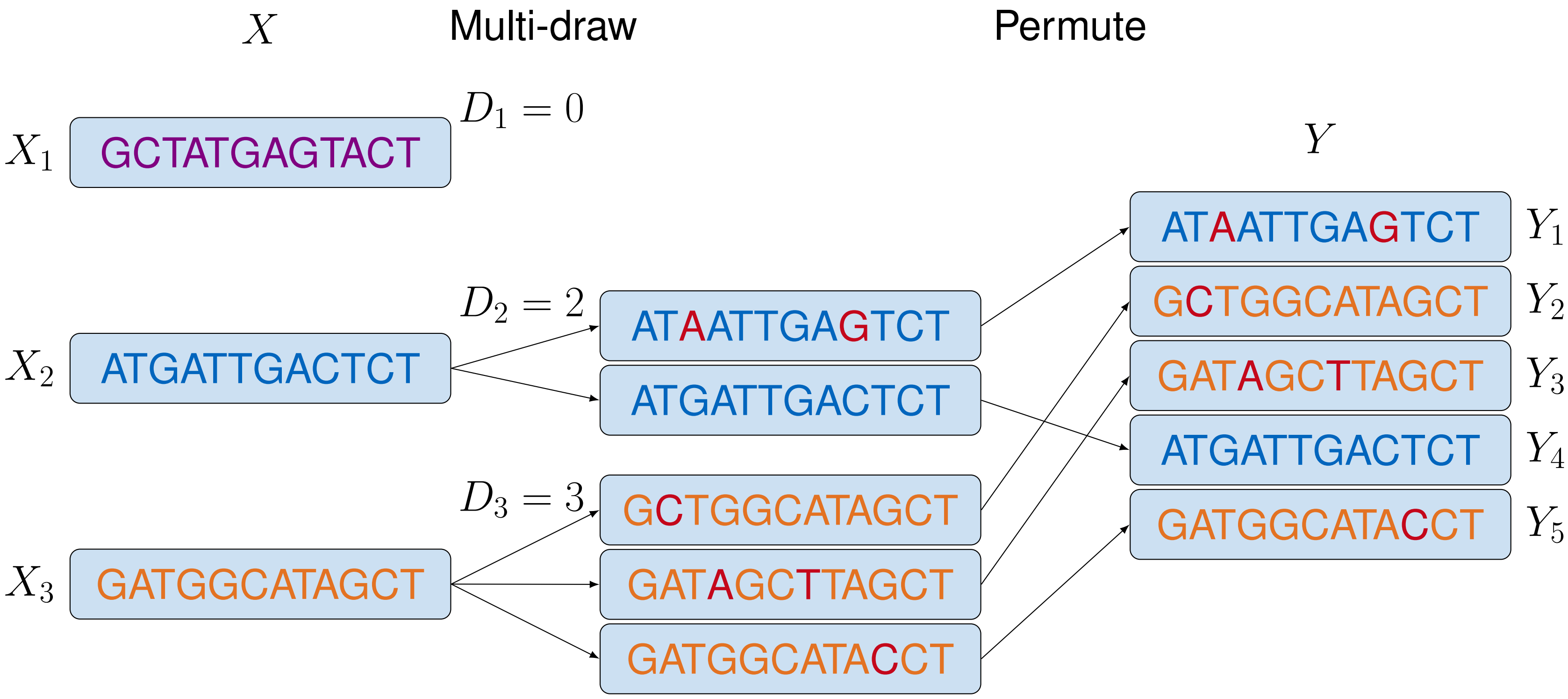


Challenges

- Draws of the multi-draw channels are random

Preliminaries - Channel Model Revisited

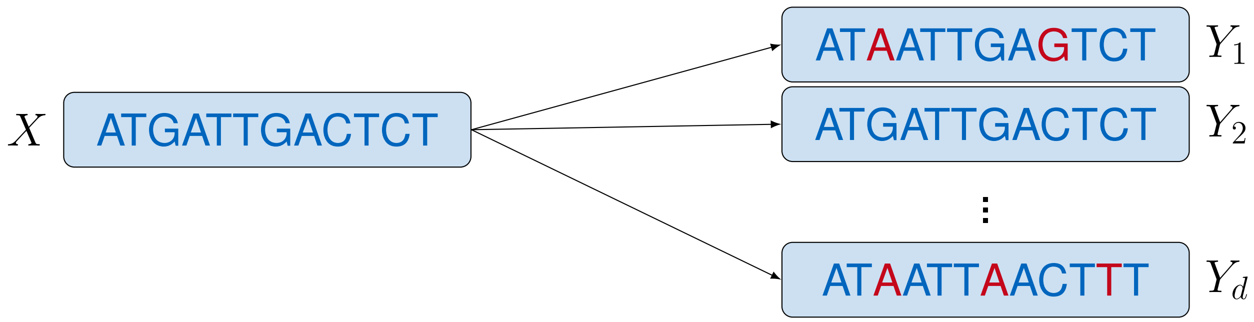
Alternative Channel Formulation



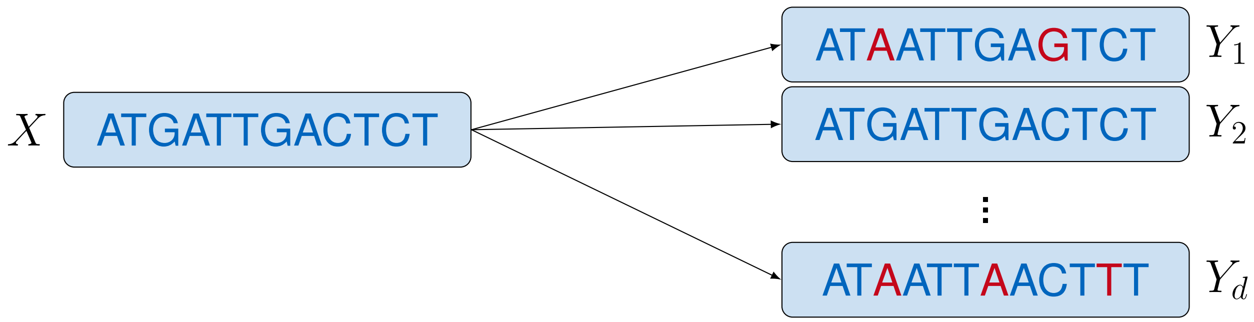
Challenges

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

Preliminaries - Multi-draw Channel [Mitzenmacher, 2006]

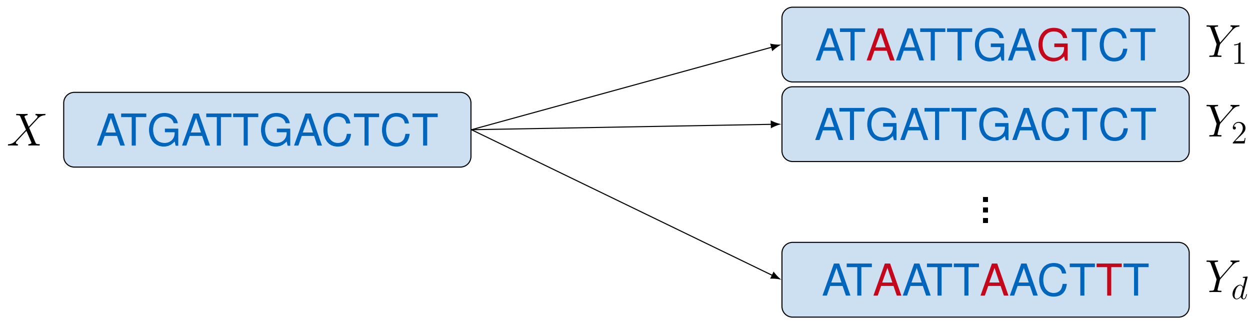


Preliminaries - Multi-draw Channel [Mitzenmacher, 2006]



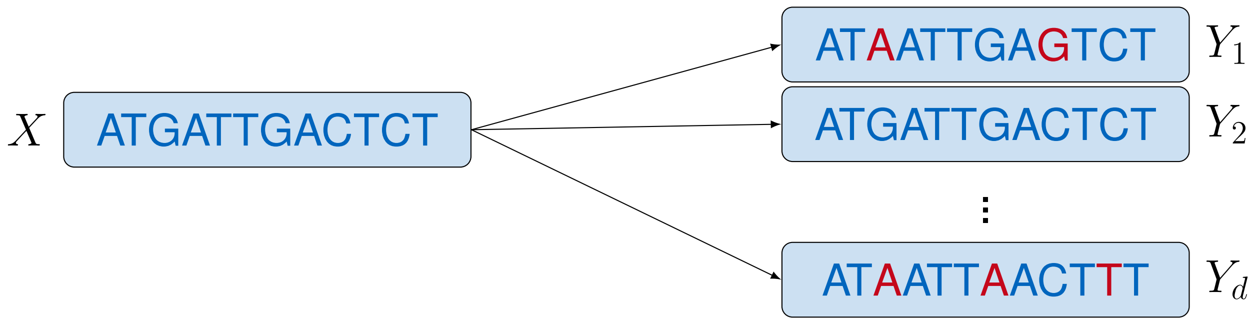
- Input: X

Preliminaries - Multi-draw Channel [Mitzenmacher, 2006]

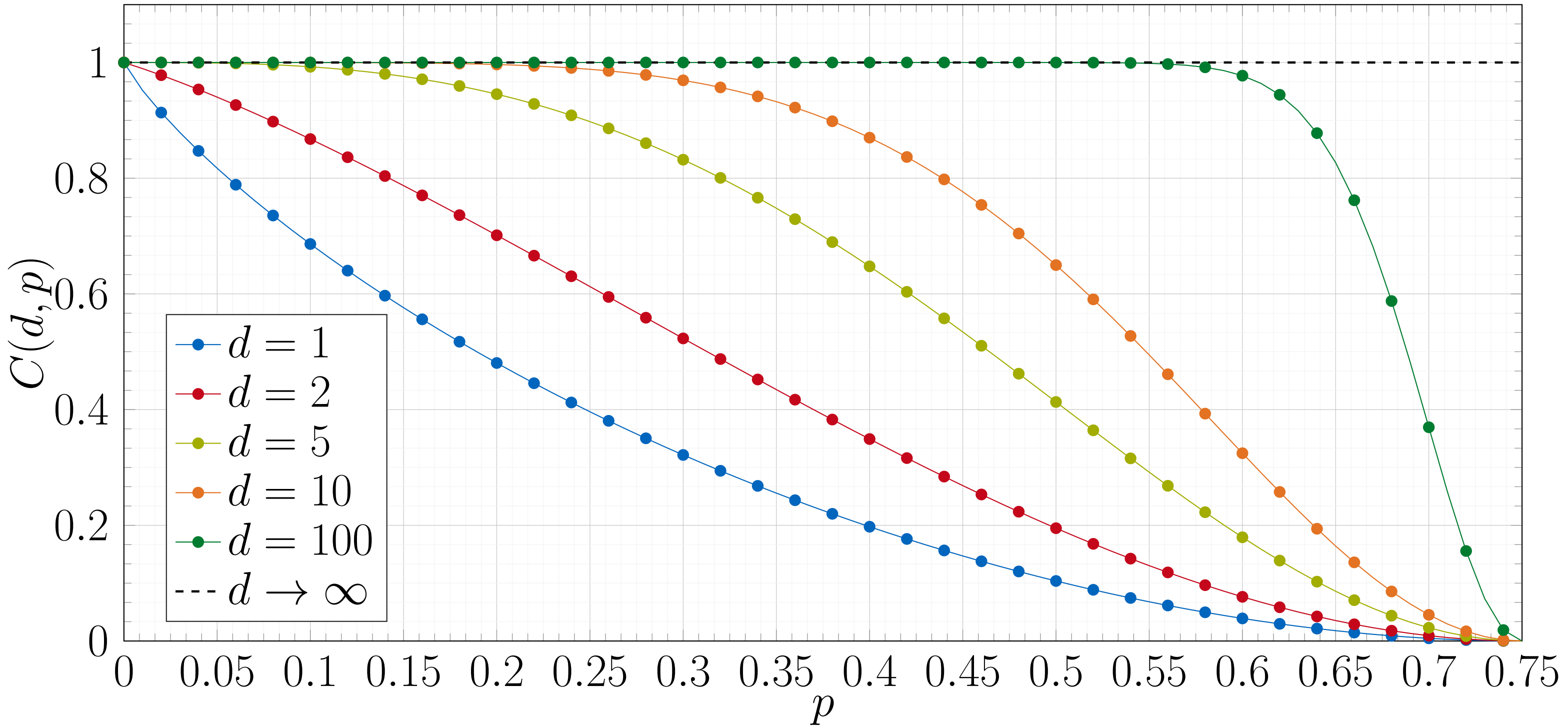


- Input: X
- Output: d output sequences $Y_1, \ldots, Y_d$
- q-ary symmetric channels: $Y_i = X + E_i$

Preliminaries - Multi-draw Channel [Mitzenmacher, 2006]



- Capacity (d draws, error probability p)



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)
- $D_i \to \text{Poi}(c)$ (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)
- $D_i \to \text{Poi}(c)$ (Poissonization)

Channel Capacity

Theorem: Channel Capacity

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

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

Channel Capacity - Parameter Range

Parameter Range

$$2\beta < 1 - H_4(2p)$$

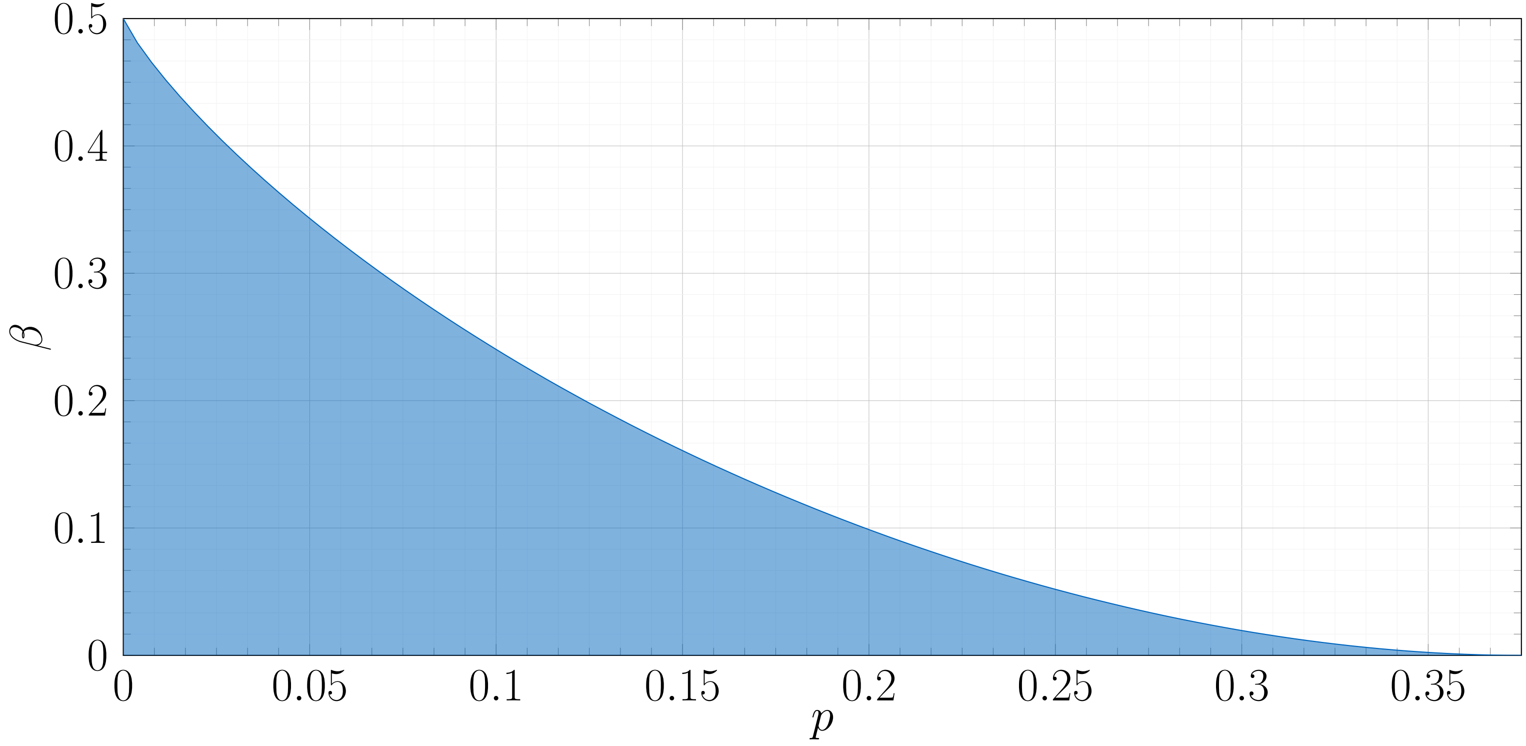
- Entropy function $H_4(p) = -(1-p)\log_4(1-p) - p\log_4\left(\frac{p}{3}\right)$

Channel Capacity - Parameter Range

Parameter Range

$$2\beta < 1 - H_4(2p)$$

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



Channel Capacity - Discussion

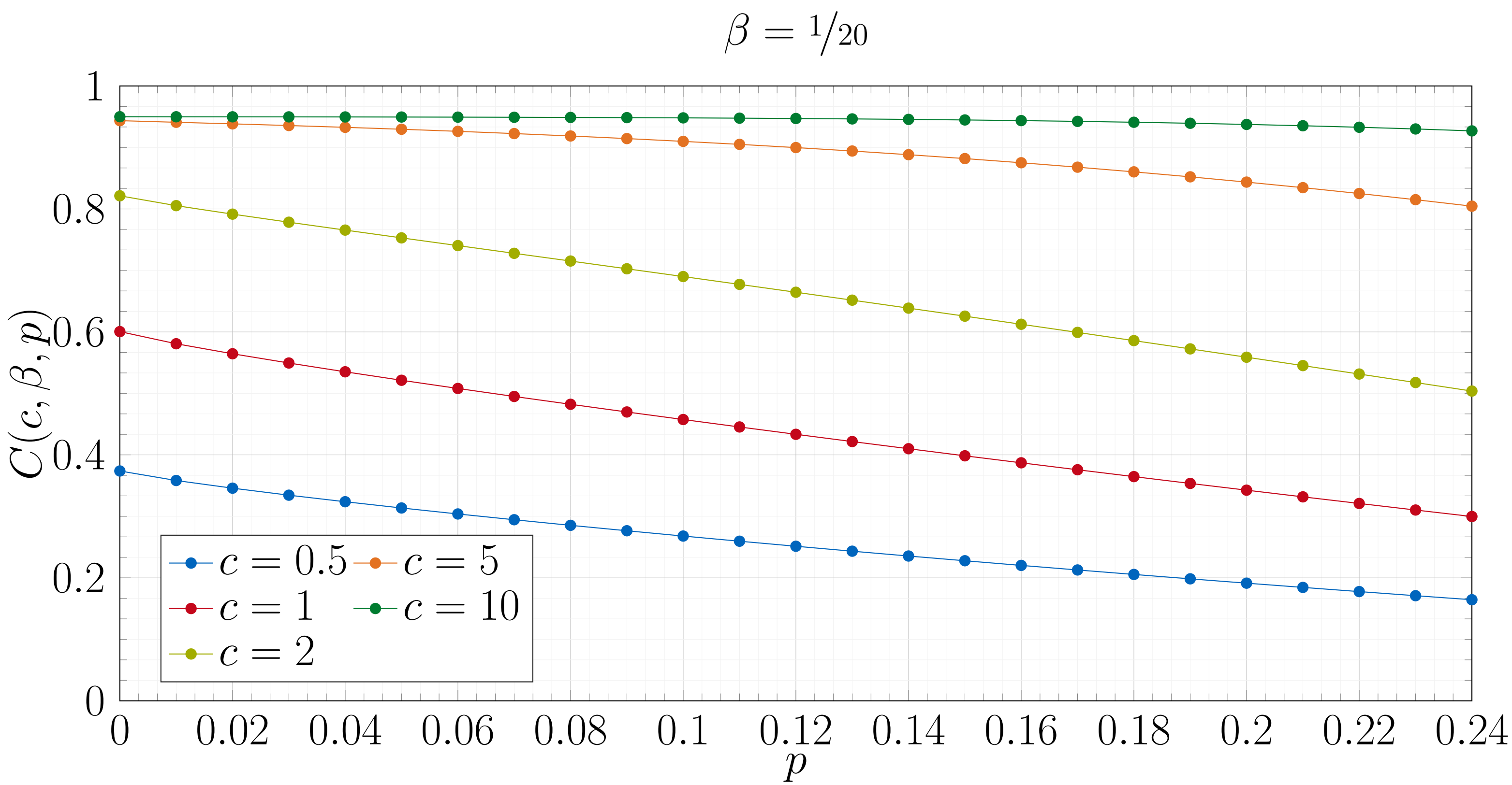
Capacity

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

Channel Capacity - Discussion

Capacity

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



Channel Capacity - Discussion

Capacity

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

Channel Capacity - Discussion

Capacity

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

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

Channel Capacity - Discussion

Capacity

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

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

$$C(c, \beta, p) = (1 - \mathrm{e}^{-c})(1 - \beta)$$

Channel Capacity - Discussion

Capacity

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

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

$$C(c,\beta,p) = (1-\mathrm{e}^{-c})(1-\beta)$$

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

Channel Capacity - Discussion

Capacity

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

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

$$C(c,\beta,p)=(1-\mathrm{e}^{-c})(1-\beta)$$

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

$$C(c,\beta,p)=1-\beta$$

Channel Capacity - Discussion

Capacity

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

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

$$C(c,\beta,p) = (1-\mathrm{e}^{-c})(1-\beta)$$

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

$$C(c,\beta,p) = 1-\beta$$

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

Channel Capacity - Discussion

Capacity

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

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

$$C(c,\beta,p) = (1-\mathrm{e}^{-c})(1-\beta)$$

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

$$C(c,\beta,p) = 1-\beta$$

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

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

Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

- Storage rate: $R_\mathsf{s} = {}^{\log_2 |\mathcal{C}|}/_{ML}$

Channel Capacity - Storage and Recovery Rate Tradeoff

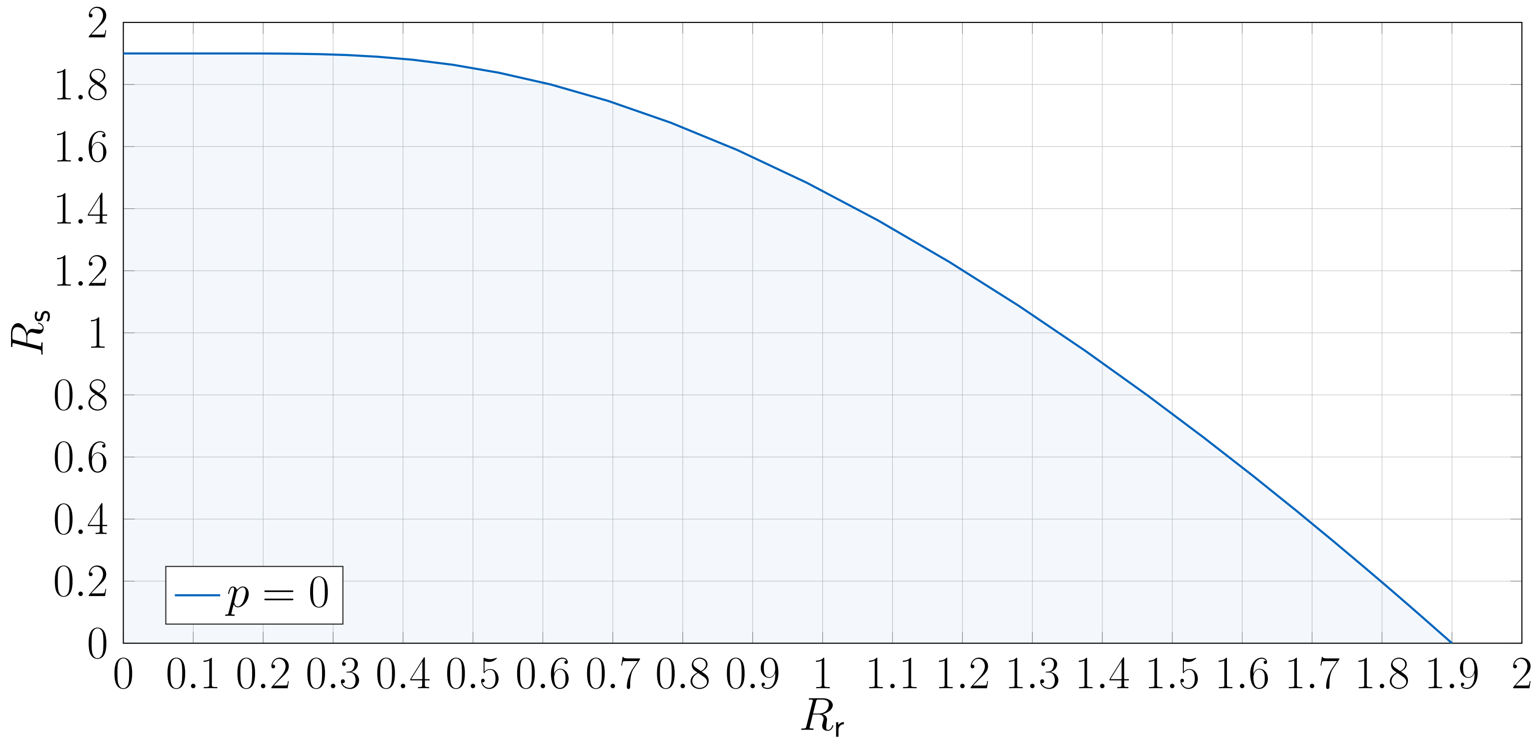
Storage and Recovery Rate Tradeoff

- Storage rate: $R_\mathsf{s} = \log_2 |\mathcal{C}| \big/ ML$
- Recovery rate: $R_\mathsf{r} = \log_2 |\mathcal{C}| \big/ NL$

Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

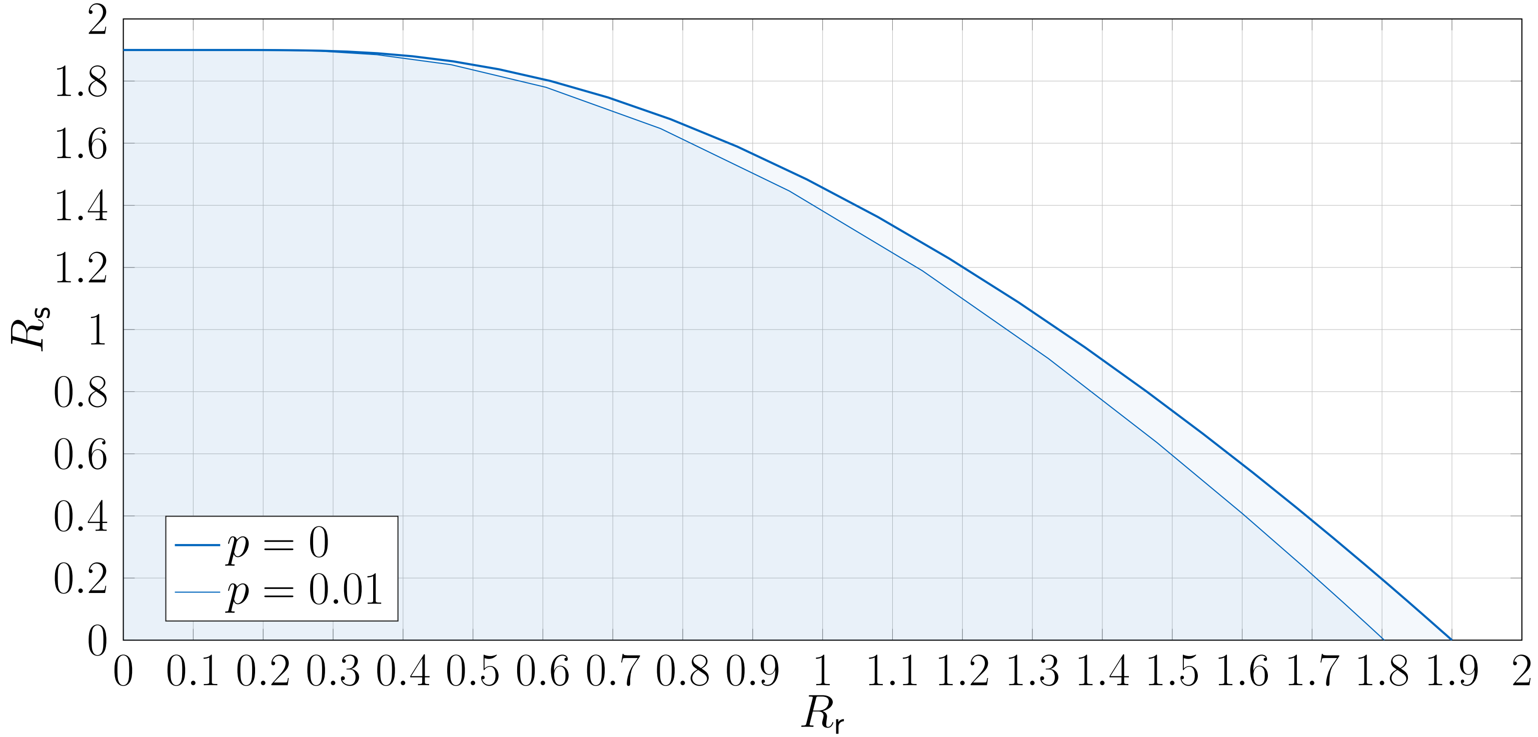
- Storage rate: $R_\mathsf{s} = {}^{\log_2 |\mathcal{C}|}/_{ML}$
- Recovery rate: $R_\mathsf{r} = {}^{\log_2 |\mathcal{C}|}/_{NL}$



Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

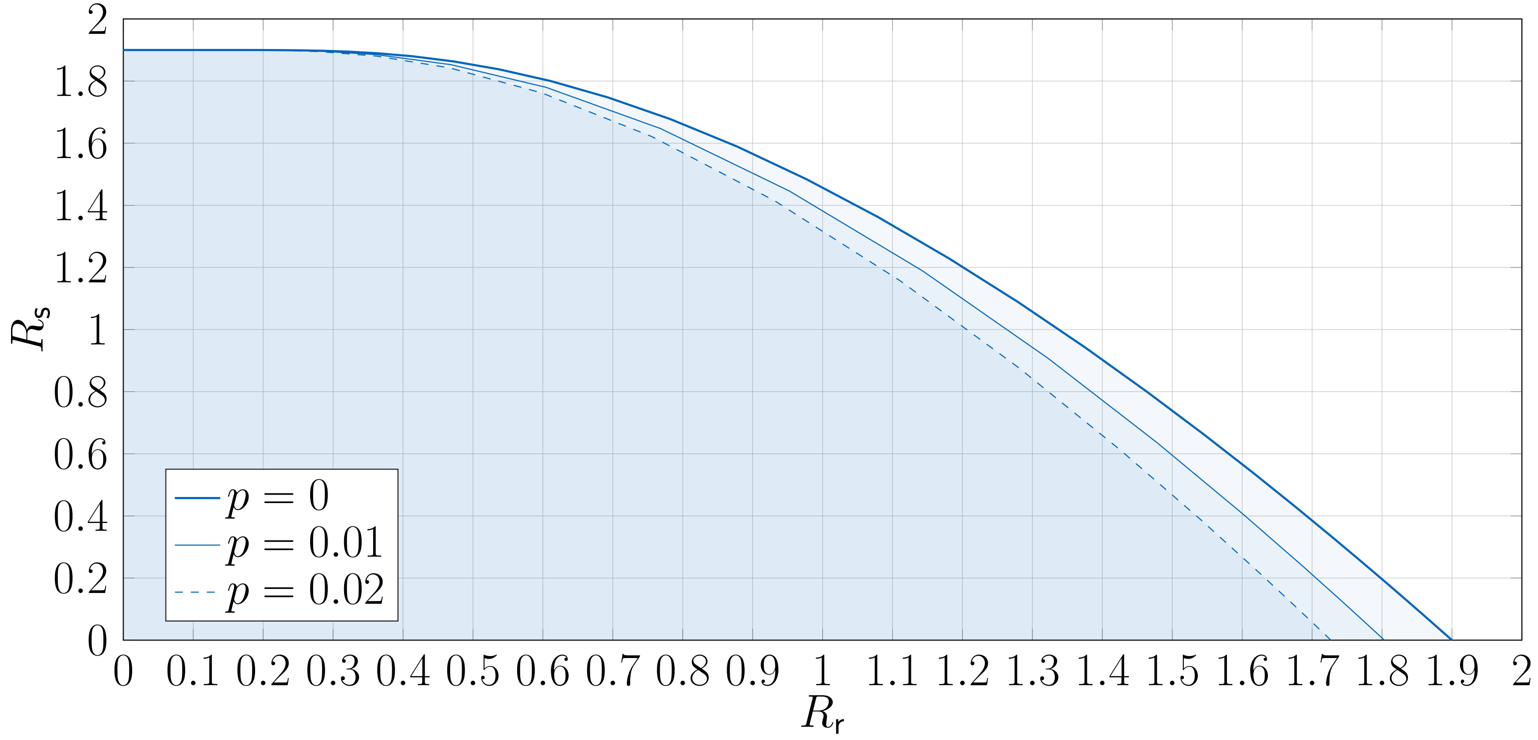
- Storage rate: $R_\mathsf{s} = {}^{\log_2 |\mathcal{C}|}/_{ML}$
- Recovery rate: $R_\mathsf{r} = {}^{\log_2 |\mathcal{C}|}/_{NL}$



Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

- Storage rate: $R_\mathsf{s} = {}^{\log_2 |\mathcal{C}|}/_{ML}$
- Recovery rate: $R_\mathsf{r} = {}^{\log_2 |\mathcal{C}|}/_{NL}$

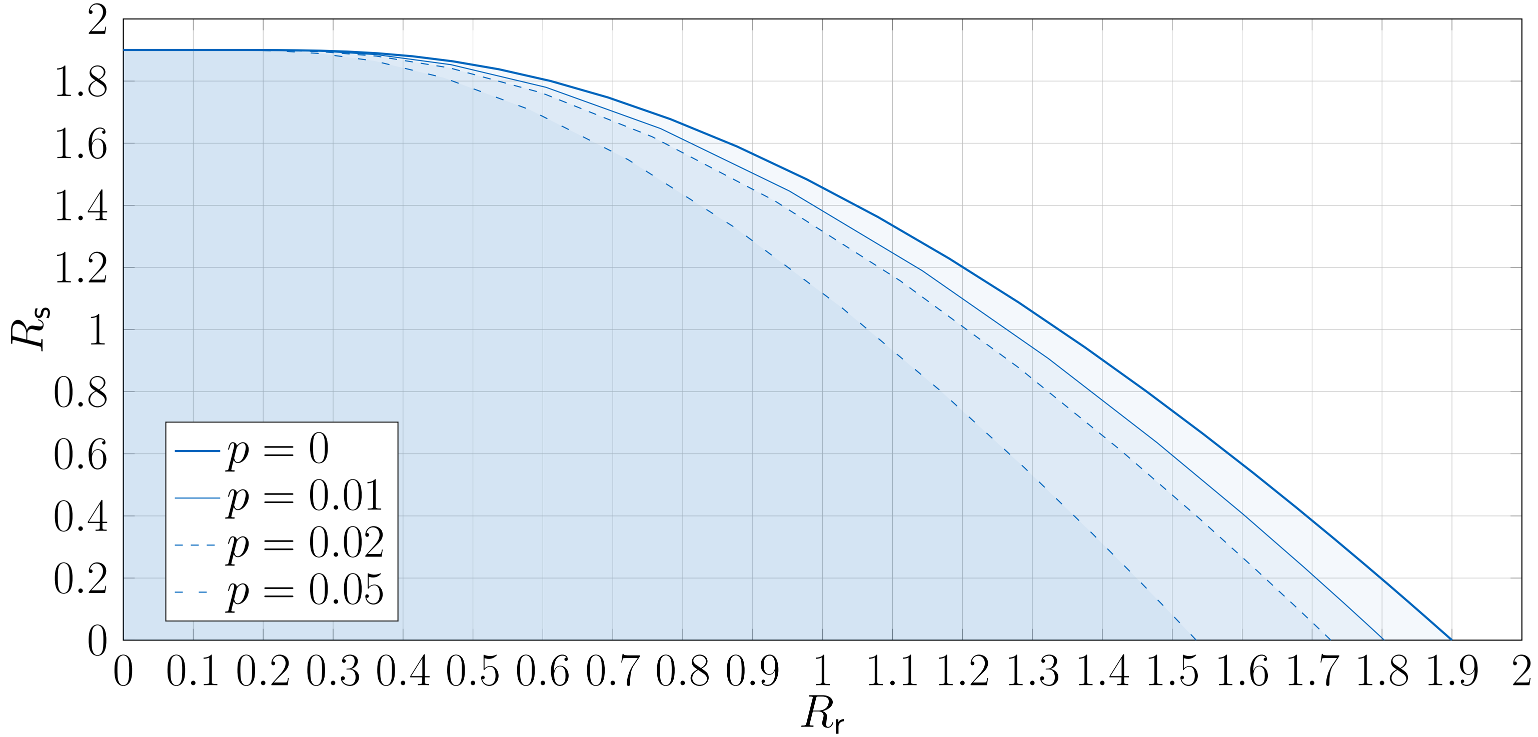


Lenz,Siegel,Wachter-Zeh,Yaakobi ”*On the Capacity of DNA-based Data Storage under Substitution Errors*”

Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

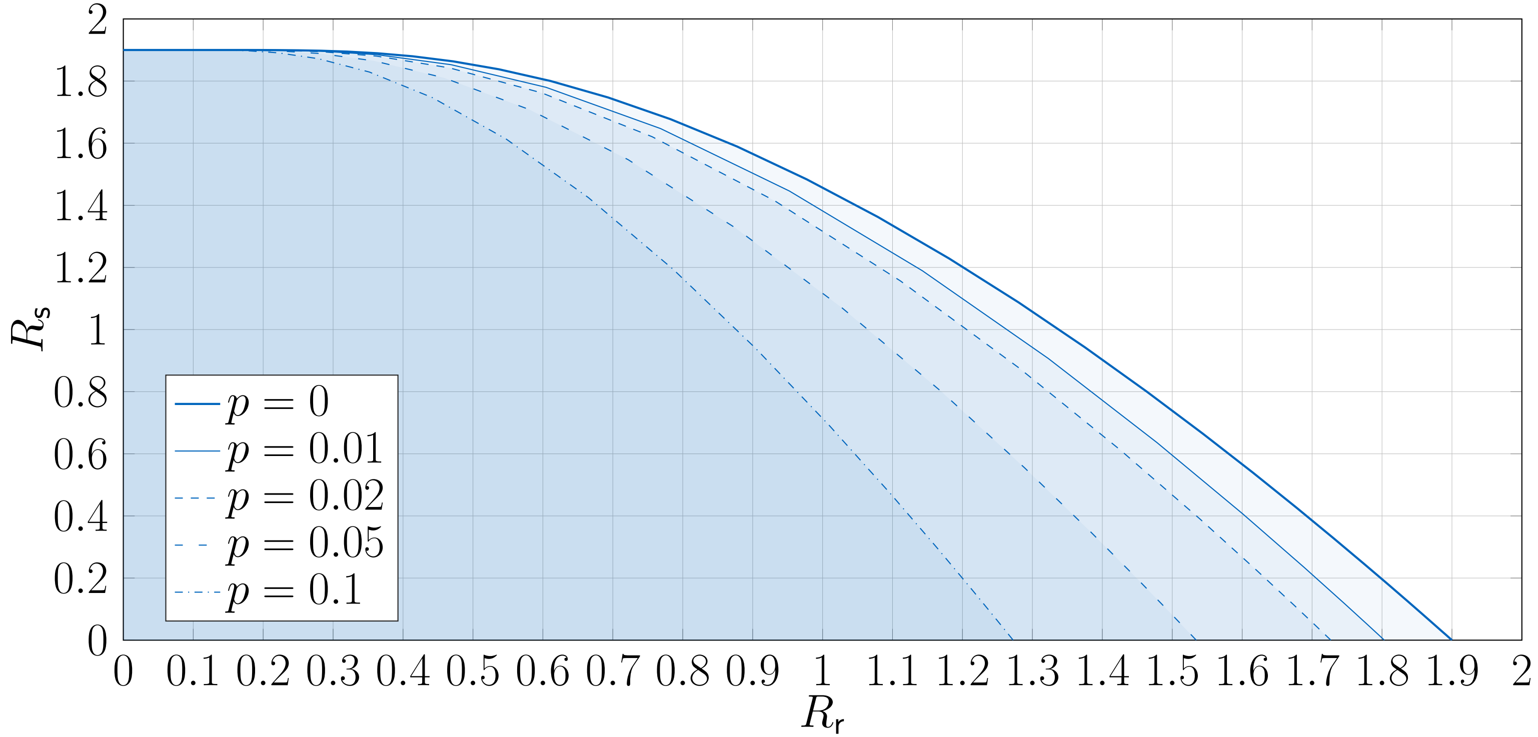
- Storage rate: $R_\mathsf{s} = {}^{\log_2 |\mathcal{C}|}/_{ML}$
- Recovery rate: $R_\mathsf{r} = {}^{\log_2 |\mathcal{C}|}/_{NL}$



Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

- Storage rate: $R_\mathsf{s} = {}^{\log_2 |\mathcal{C}|}/_{ML}$
- Recovery rate: $R_\mathsf{r} = {}^{\log_2 |\mathcal{C}|}/_{NL}$

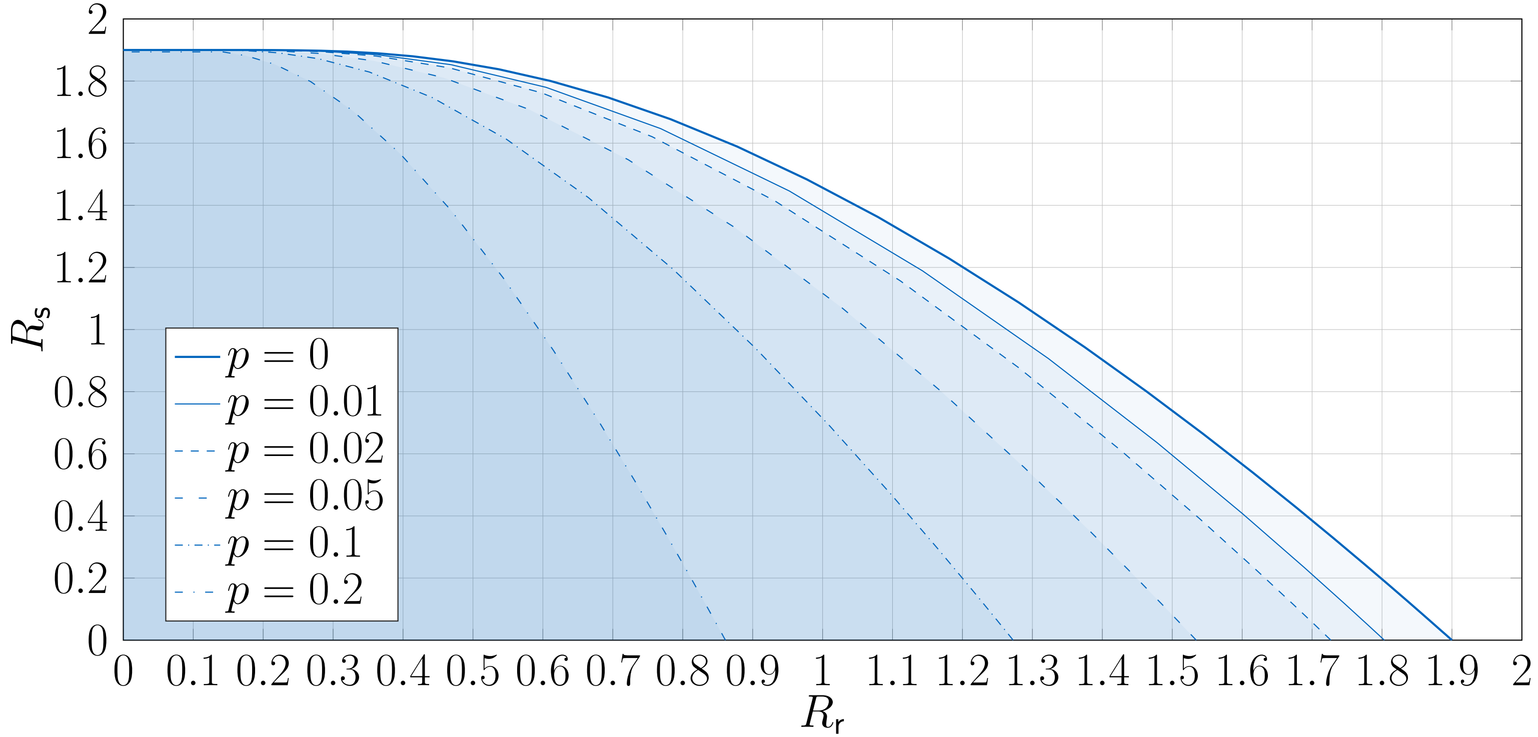


Lenz,Siegel,Wachter-Zeh,Yaakobi "*On the Capacity of DNA-based Data Storage under Substitution Errors*"

Channel Capacity - Storage and Recovery Rate Tradeoff

Storage and Recovery Rate Tradeoff

- Storage rate: $R_\mathsf{s} = {}^{\log_2 |\mathcal{C}|}/_{ML}$
- Recovery rate: $R_\mathsf{r} = {}^{\log_2 |\mathcal{C}|}/_{NL}$



Summary & Outlook

Summary

- Capacity of the DNA storage channel under substitution errors

Summary & Outlook

Summary

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

Summary & Outlook

Summary

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

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!