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

Andreas Lenz*, Paul H. Siegel†, Antonia Wachter-Zeh*, and Eitan Yaakobi‡
*Institute for Communications Engineering, Technical University of Munich, Germany D-80333
†Department of Electrical and Computer Engineering, University of California, San Diego, California
‡Computer Science Department, Technion – Israel Institute of Technology, Haifa, Israel
Emails: andreas.lenz@mytum.de, psiegel@ucsd.edu, antonia.wachter-zeh@tum.de, yaakobi@cs.technion.ac.il

***Abstract*—Advances in biochemical technologies, such as synthesizing and sequencing devices, have fueled many recent experiments on archival digital data storage using DNA. In this paper we study the information-theoretic capacity of such storage systems. The channel model incorporates the main properties of DNA-based data storage. We present the capacity of this channel for the case of substitution errors inside the sequences and provide an intuitive interpretation of the capacity formula for relevant channel parameters. We compare the capacity to rates achievable with a sub-optimal decoding method and conclude with a discussion on cost-efficient DNA archive design.**

I. Introduction

DNA-based data storage is a novel approach for long-term archiving of digital data. It has drawn recent attention due to significant advances in biochemical technologies, such as synthesizing and sequencing of DNA. Experiments addressing many different aspects of digital data storage, such as reliability, lifetime, random-access, and efficiency have been published in the last decade. At the same time, the unique nature of DNA-based storage systems has fueled theoretical investigations in computational biology, coding theory, information theory, and signal processing.

The process of writing and reading digital data in DNA-based data storage basically involves three main steps. First, the digital binary data is encoded into many short vectors over the alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$, which are then synthesized as DNA strands. In most experiments, each strand is synthesized many times such that multiple copies of each strand are present. Second, those strands are transferred into a storage medium that preserves the chemical structure of DNA and ensures robustness over a long period of time. Third and finally, when accessing the data inside the archive, the DNA strands from the storage medium are sequenced. Due to the nature of the sequencer, it is generally not possible to choose which strands are sequenced. (In contrast to [1], [2], we study the raw system that does not include the use of primers appended to the DNA strand to allow sequencing specific strands.) The synthesis and sequencing may induce insertion and deletion errors, as well as substitution errors, in the DNA strands. Using the sequencing data, a decoder then estimates the original digital data.

As a step toward analyzing the information-theoretic capacity of DNA-based storage, this work considers the *noisy drawing channel* that models the pipeline from synthesized to sequenced DNA strands. It incorporates the unordered nature of the sequencing process by modeling the received strands as random draws of the input sequences together with substitution errors inside the DNA strands. Prior work [3] has discussed this channel for the noiseless case. We review our recent results about the capacity of the noisy drawing channel, reported in [4], and give an intuitive explanation of the capacity formula.

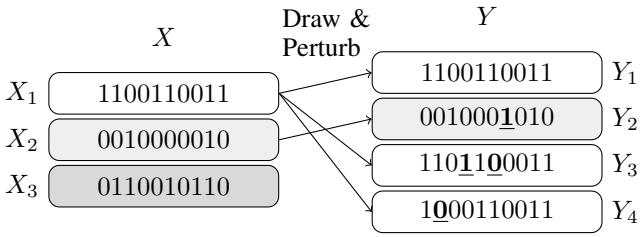

Fig. 1: Exemplary realization of the DNA storage channel with $M = 3$ and $N = 4$. Shades highlight the origin of the received sequences. This origin is however *unknown* to the receiver.

(Our work [4] has recently been extended to asymmetric error channels [5].) We further examine achievable rates with a sub-optimal decoding method. Finally, using the capacity formula, we present an optimization problem that allows to design cost-efficient storage systems. This work has been published in [6].

II. DNA Storage Channel Model

A. The Noisy Drawing Channel

The input of the DNA storage channel is M sequences $X_1, \dots, X_M$ where each X_i is a vector of length L over the binary alphabet $\Sigma = \{0, 1\}$. (Our results can be extended to the quaternary DNA alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$.) From these input sequences, a total of $N = cM$ sequences are drawn with replacement, each uniformly at random, and received with errors. The parameter $c > 0$ represents the *coverage* of the draws. The output of the channel is then given by N sequences $Y_1, \dots, Y_N$, each of length L. Each sequence Y_j is obtained by drawing a random input sequence X_{I_j} and transmitting it over a binary symmetric channel (BSC) with error probability p. We can think of the input and output sequences as matrices $X = (X_1, \dots, X_M) \in \Sigma^{M \times L}$ and $Y = (Y_1, \dots, Y_N) \in \Sigma^{N \times L}$. Fig. 1 illustrates an exemplary realization of this channel.

B. Multidraw Channel

An important component of the noisy drawing channel is the so-called *multidraw* or *binomial* channel. It captures the fact that each input sequence X_i is observed through D_i noisy output sequences, each originating from the same input sequence. The multidraw channel is parameterized by the number of draws $d \in \mathbb{N}$ and error probability $0 \leq p \leq 1$. Its capacity, derived in [7], is as follows.

$$C_d = 1 + \sum_{k=0}^{d} \binom{d}{k} p^k (1-p)^{d-k} \log \frac{1}{1 + p^{d-2k}(1-p)^{2k-d}}.$$

III. Capacity of the Noisy Drawing Channel

Since the channel input is M sequences, each of length L, a code over the noisy drawing channel is a set $\mathcal{C} \subseteq \Sigma^{M \times L}$.

Its *storage rate* is defined to be the number of bits that can be stored per nucleotide, i.e., $R_{\mathsf{s}} = \log|\mathcal{C}|/ML$. Similarly we can define the *recovery rate* of a code $\mathcal{C}$ as the number of information bits that can be retrieved per nucleotide that is sequenced, i.e., $R_{\mathsf{r}} = \log|\mathcal{C}|/NL$.

Assume a codeword $X \in \mathcal{C}$ has been transmitted over the noisy drawing channel and $Y \in \Sigma^{N\times L}$ has been received. A decoder for a code $\mathcal{C}$ is then a mapping $\mathsf{dec} : \Sigma^{N\times L} \mapsto \mathcal{C} \cup \{\mathsf{fail}\}$, $\mathsf{dec}(Y) = \hat{X}$, where fail denotes a decoding failure, i.e., the decoder cannot find any suitable codeword.

We will set $M = 2^{\beta L}$ and $N = cM$ for some fixed $0 < \beta < 1, 0 < c$, and let L go to infinity. Let β, c, and p be fixed and given. We say a rate R_{s} is *achievable*, if there exists a family of codes $\mathcal{C}(M \times L) \subseteq \Sigma^{M\times L}$ with storage rate R_{s} together with a decoder $\mathsf{dec} : \Sigma^{N\times L} \mapsto \mathcal{C}(M \times L) \cup \{\mathsf{fail}\}$ such that the decoding error probability tends to zero, as $L \to \infty$, with $M = 2^{\beta L}$ and $N = cM$. The capacity of the noisy drawing channei, $C(\beta, c, p)$, is the supremum of achievable rates and is given as follows [4].

Theorem 1 *Fix $0 < c$, $0 \le p < \frac{1}{8}$, and $0 < \beta < \frac{1-H(4p)}{2}$. Then, the capacity of the noisy drawing channel is given by*

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

where $\mathsf{Poi}_c(d) = \frac{e^{-c}c^d}{d!}$ is the probability mass function of the Poisson distribution with expected value c and $H(p)$ is the binary entropy function.

The result holds for, e.g., all $p \le 0.075$ and $\beta \le \frac{1}{20}$. Most recent experiments have parameters within this region [8], [9].

Conceptually, the noisy drawing channel can be split into two sub-channels. The first sub-channel transmits each input sequence $X_i,\ i = 1, \ldots, M$, over one of M parallel multidraw channels, each with D_i draws. The second sub-channel then randomly permutes the resulting set of sequences comprising the draws of all $X_i,\ i = 1, \ldots, M$. The first term of the capacity formula corresponds to the capacity of the first sub-channel. The rate loss associated with the uncertainty introduced by the second sub-channel is exactly the second term in the formula. A precise analysis supports this interpretation [4], [10].

Fig. 2 shows the capacity for $\beta = 1/20$ and different values of coverage c over a range of values of p. Note that the plot is limited to error probabilities of at most $p = 0.075$ due to the parameter limitation in Theorem 1. The figure also shows achievable rates using a suboptimal decoder based on majority voting. This decoder first performs a bitwise majority decision on all sequences that stem from the same input sequence. We see that the overall rate loss with respect to the capacity is relatively small over a range of channel parameters.

IV. Design of Cost-Efficient DNA Archives

Most publications to date focus on the storage rate R_{s} to evaluate their results. More recently, however, the interest in efficient design with respect to both storage rate R_{s} and recovery rate R_{r} has increased. We now present an optimization-based approach to designing cost-efficient DNA storage systems. We let β and p be fixed parameters to be chosen by the system engineer. While β is usually determined by the length of the DNA sequences and the amount of digital data to be stored, p is given by the synthesis and sequencing technologies. The costs associated with DNA-based data storage are mainly due

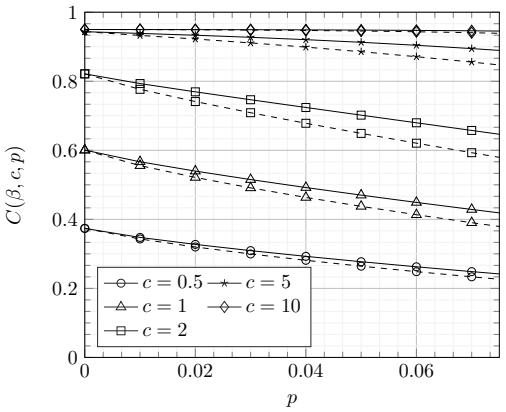

Fig. 2: Capacity of the noisy drawing channel for different values of c, over a range of error probabilities p, with $\beta = \frac{1}{20}$. The dashed lines show achievable rates for suboptimal majority vote decoding.

to the synthesis and sequencing of DNA strands. To this end, assume we are given a synthesis machine that incurs a cost of γ_{s} per nucleotide. Further, we use a sequencing machine that has an associated cost γ_{r} per read of a single nucleotide of DNA. Using a code of storage rate R_{s} and recovery rate R_{r}, the total cost associated with writing and reading a single bit to and from the archive is

$$\gamma(\beta, c, p) = \frac{\gamma_{\mathsf{s}}}{R_{\mathsf{s}}} + \frac{\gamma_{\mathsf{r}}}{R_{\mathsf{r}}} = \frac{1}{C(\beta, c, p)} (\gamma_{\mathsf{s}} + c\gamma_{\mathsf{r}}) \tag{2}$$

where we assumed the usage of a capacity-achieving storage code, i.e., $R_{\mathsf{s}} = C(\beta, c, p)$, and used the relation $R_{\mathsf{s}} = cR_{\mathsf{r}}$.

Note that in comparison to [3] we additionally incorporate the error probability p into the system design. We can optimize (2) over c for given β and p. Currently the synthesis cost is a factor of roughly 10^4 larger than the sequencing cost [3] and we thus set $\frac{\gamma_{\mathsf{s}}}{\gamma_{\mathsf{r}}} = 10^4$. For $p = 0.02$ and $\beta = \frac{1}{20}$, one obtains that $c^* \approx 11.4$ minimizes the cost, while for $p = 0.05$, we obtain $c^* \approx 14$. Note that smaller synthesis costs will push the optimum c^* towards smaller values, since the sequencing costs become more apparent. One can extend this cost optimization to reflect the dependence of costs on the synthesis and sequencing quality p and perform a joint optimization over c and p.

References

[1] S. M. H. T. Yazdi *et al.*, "A rewritable, random-access DNA-based storage system," *Sci. Rep.*, vol. 5, no. 14138, Nov. 2015.

[2] S. M. H. T. Yazdi *et al.*, "Portable and error-free DNA-based data storage," *Sci. Rep.*, vol. 7, no. 5011, Dec. 2017.

[3] R. Heckel *et al.*, "Fundamental limits of DNA storage systems," *Proc. Int. Symp. Inf. Theory*, Jun. 2017, pp. 3130-3134.

[4] A. Lenz *et al.*, "Achieving the capacity of the DNA storage channel," *Proc. Int. Conf. Acoust., Speech, Sig. Process.*, May 2020, pp. 8846-8850.

[5] N. Weinberger and N. Merhav, "The DNA Storage Channel: Capacity and Error Probability," arXiv:2109.12549 [cs, math], Sep. 2021.

[6] A. Lenz *et al.*, "On the capacity of DNA-based data storage under substitution errors," *Proc. Int. Conf. Visual Commun. Image Proc.*, Dec. 2021, pp. 1-5.

[7] M. Mitzenmacher, "On the theory and practice of data recovery with multiple versions," *Proc. Int. Symp. Inf. Theory.*, Jul. 2006, pp. 982-986.

[8] R. Heckel *et al.*, "A characterization of the DNA data storage channel," *Sci. Rep.*, vol. 9, no. 9663, Jul. 2019.

[9] A. Lenz *et al.*, "Coding over sets for DNA storage," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2331–2351, Apr. 2020.

[10] A. Lenz *et al.*, "An upper bound on the capacity of the DNA storage channel," *Proc. Inf. Theory Workshop*, Aug. 2019, pp. 1-5.