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

(Invited Paper)

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

2021 International Conference on Visual Communications and Image Processing (VCIP) | 978-1-7281-8551-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/VCIP53242.2021.9675410

***Abstract*—Advances in biochemical technologies, such as synthesizing and sequencing devices, have fueled manifold recent experiments on archival digital data storage using DNA. In this paper we review and analyze recent results on information-theoretic aspects of such storage systems. The discussion focuses on a channel model that incorporates the main properties of DNA-based data storage. Namely, the user data is synthesized many times onto a large number of short-length DNA strands. The receiver then draws strands from the stored sequences in an uncontrollable manner. Since the synthesis and sequencing are prone to errors, a received sequence can differ from its original strand, and their relationship is described by a probabilistic channel. Recently, the capacity of this channel was derived for the case of substitution errors inside the sequences. We review the main techniques used to prove a coding theorem and its converse, showing the achievability of the capacity and the fact that it cannot be exceeded. We further provide an intuitive interpretation of the capacity formula for relevant channel parameters, compare with sub-optimal decoding methods, and conclude with a discussion on cost-efficiency.**

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. Manifold experiments [1]–[6] have been published in the last decade, addressing many different aspects of digital data storage, such as reliability, lifetime, random-access, and efficiency. At the same time, the unique nature of DNA-based storage systems has fueled theoretical investigations inside a variety of research fields, such as computational biology, coding theory, information theory and signal processing. This uniqueness is mainly due to the technologies used to synthesize and sequence DNA.

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, this is often an uncontrollable procedure in the sense that it is not possible to choose which strands are sequenced.[1] Using the sequencing data, a decoder then estimates the original digital data. There are several aspects that distinguish DNA-based storage systems from conventional transmission or storage systems. First, the unordered nature of sequencing is seldom observed in traditional communication systems. Next, a DNA strand is a complex molecule where different nucleotides interact with one another, requiring the consideration of its stability when designing the DNA sequences. Finally, the synthesis and sequencing are prone to insertion and deletion errors, which are errors observed in few other communication systems.

Most information-theoretic studies related to DNA-based data storage discuss insertion and deletion error correction. Classical papers on this topic are those by Gallager [7] and Davey and Mackay [8]. More recently, increased interest towards channel models with multiple transmissions over a channel impaired by insertion and deletion errors arose, due to the existence of multiple copies of each stored strand in DNA-based storage systems. For example, [9], [10] study sequence reconstruction from multiple DNA sequences, whereas [11] discusses the case, where the case of coded original sequences. Decoding algorithms and achievable information rates for a channel whose output comprises several sequences obtained over independent insertion and deletion channels are discussed in [12], [13]. A related channel featuring multiple sequences with errors and random shuffling has been discussed in [14].

This work deals with the so-called *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 [15] has discussed this channel for the noiseless case. We review recent results about the

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 801434). This work was also supported by NSF Grant CCF-BSF-1619053 and by the United States-Israel BSF grant 2015816.

[1]There are studies [3], [4] that have developed methods for random access and for sequencing of specific strands. This was accomplished by designing primers that are appended to the DNA strand. Here however, we are studying the *raw* system without the usage of such additional primers.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 09,2023 at 03:21:26 UTC from IEEE Xplore. Restrictions apply.

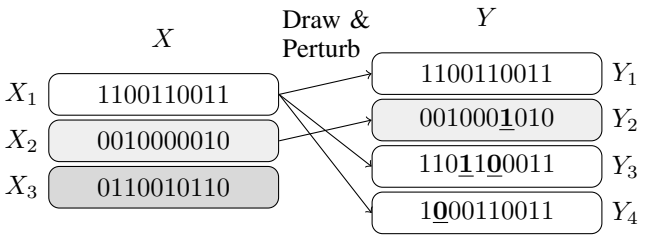

Fig. 1. Exemplary realization of the DNA storage channel with $M = 3$ and $N = 4$. Background colors highlight the origin of the received sequences.

capacity of the noisy drawing channel, reported in [16]. In particular, we give an intuitive explanation of the capacity formula and a coarse overview over the ingredients required to prove these capacity results. We further compare achievable rates with sub-optimal decoding methods with the capacity. Finally, using the capacity formula, we present an optimization problem that allows to design cost-efficient storage systems. Note that recently, an extension of our work [16] to a broader class of constituent channels has been published in [17].

II. Preliminaries

A. The Noisy Drawing Channel

For an integer n, we denote by $[n] = \{1, 2, \ldots, n\}$ the set of positive integers up to n. The input of the DNA storage channel is M sequences $X_1, \ldots, X_M$ where each $X_i = (X_{i,1}, \ldots, X_{i,L}) \in \Sigma^L$, $i \in [M]$, is a vector of length L over the binary alphabet $\Sigma = \{0, 1\}$. From these input sequences, a total of N sequences are drawn with replacement, each uniformly at random, and received with errors. Denote by $I_j \in [M]$ the index of the j-th drawn input sequence. We assume that the draws I_j are i.i.d. uniform random variables with $\mathbb{P}(I_j = i) = \frac{1}{M}$ for all $j \in [N]$ and $i \in [M]$. The output of the channel is then given by N sequences $Y_j = (Y_{j,1}, \ldots, Y_{j,L}) \in \Sigma^L$, $j \in [N]$, 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 crossover probability p. That is, the individual output symbols are given by

$$Y_{j,k} = \begin{cases} X_{I_j,k} & \text{with probability } 1-p, \\ 1 - X_{I_j,k} & \text{with probability } p \end{cases},$$

for all $j \in [N], k \in [L]$. For convenience, we stack all input and output sequences to matrices $X = (X_1, \ldots, X_M) \in \Sigma^{M \times L}$ and $Y = (Y_1, \ldots, Y_N) \in \Sigma^{N \times L}$, such that each sequence is a row of the corresponding matrix. Figure 1 illustrates an exemplary realization of this channel.

We continue by defining important random variables associated with this channel model. Throughout the paper we will use the random variables $D_i = |\{j \in [N] : I_j = i\}|$, $i \in [M]$, which count the number of times the i-th input sequence has been drawn and $Q_d = |\{i \in [M] : D_i = d\}|$, $d = 0, \ldots, N$, that denote the number of input sequences that have been drawn a total of d times.

B. Multidraw Channel

An important component of the noisy drawing channel is the so-called *multidraw* or *binomial* channel. Its relevance is due to the fact that each input sequence X_i is observed through D_i output sequences, each originating from the same input sequence. The multidraw channel has first been discussed in [18], where the capacity was derived together with considerations about practical transmission over this channel. The multidraw channel is parameterized by the number of draws $d \in \mathbb{N}$ and error probability $0 \leq p \leq 1$. It is a discrete memoryless channel with binary input and d binary output symbols, where each symbol is equal to the input with probability $1 - p$ and different from the input with probability p. The channel output is therefore obtained by a d-fold repeated transmission over a standard binary symmetric channel. The capacity of the multidraw channel is directly obtained using the fact that it is a discrete memoryless channel and computes to

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

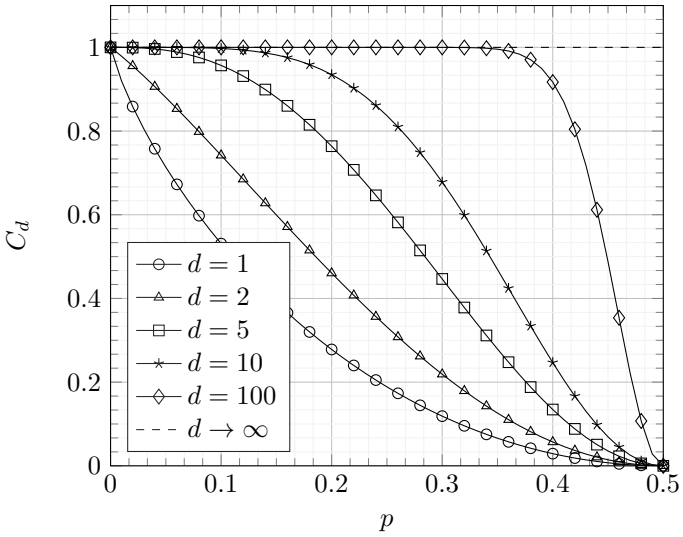

Fig. 2. Capacity of the multidraw channel for different number of draws d, over error probabilities p

For an illustration of the capacity for different number of draws, see Figure 2. Note that $C_0 = 0$, $C_1 = 1 + p \log p + (1-p)\log(1-p)$ and $C_0 \leq C_1 \leq C_2 \leq C_3 \leq \ldots$, where the inequalities are due to the data processing inequality.

III. Capacity of the Noisy Drawing Channel

Before we present the results on the capacity of the noisy drawing channel, we formally define achievable information rates and the associated capacity. We start with defining error-correcting codes over the presented 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.,

$$R_{\mathsf{s}} = \frac{\log |\mathcal{C}|}{ML}. \tag{1}$$

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.,

$$R_{\mathsf{r}} = \frac{\log |\mathcal{C}|}{NL}. \tag{2}$$

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 09,2023 at 03:21:26 UTC from IEEE Xplore. Restrictions apply.

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. A *decoding success* is the event $\mathsf{dec}(Y) = X$ and, conversely, the complementary event is a *decoding error*. Consequently, assuming that a codeword $X \in \mathcal{C}$, chosen uniformly at random, is transmitted over the channel, the (average) error probability of a code $\mathcal{C}$ with a decoder dec is the probability

$$\mathbb{P}(\mathcal{E}) = \frac{1}{|\mathcal{C}|} \sum_{X\in\mathcal{C}} \mathbb{P}(\mathsf{dec}(Y) \neq X),$$

where Y is the result of transmitting X over the noisy drawing channel. For our asymptotic statements in this paper we will set $M = 2^{\beta L}$ and $N = cM$ for some fixed $0 < \beta < 1, 0 < c$, and let L go to infinity. With this relationship we arrive at the following definition of achievable rates. 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 $\mathbb{P}(\mathcal{E}) \to 0$ tends to zero, as $L \to \infty$, with $M = 2^{\beta L}$ and $N = cM$. We can use this definition to define the capacity of the noisy drawing channel as the supremum of achievable rates, i.e., $C(\beta, c, p) = \sup\{R_{\mathsf{s}} : R_{\mathsf{s}} \text{ is achievable}\}$. By this definition, the capacity is a function that exclusively depends on the channel parameters β, c, and p. Explicitly, the capacity is given as follows [16].

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

$$C(\beta, c, p) = \sum_{d=0}^{\infty} \mathsf{Poi}_c(d) C_d - \beta(1 - \mathrm{e}^{-c}), \tag{3}$$

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

Note that this theorem holds only for relatively small noise values $p < \frac{1}{8}$ and moderate number of sequences β and a valid capacity expression remains unknown for a larger range of parameters. However, the result holds for, e.g., all $p \leq 0.075$ and $\beta \leq \frac{1}{20}$. Notably, most current experiments report parameters within this region [19], [20].

A. Interpretation and Intuition

We now proceed with giving an intuitive explanation of the capacity formula. Conceptually, the noisy drawing channel can be split into two sub-channels, for an illustration see Figure 3. The first sub-channel transmits each input sequence X_i, $i \in [M]$ 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, $i \in [M]$. It is easy to check that the input-output relation is the same as in the original model. The capacity of the first sub-channel is obtained as follows. First note that the random variables D_i individually converge to Poisson distributions with mean c as $L \to \infty$ since the D_i's are binomial distributed with $N = cM$ trials and success probability $\frac{1}{M}$. While the variables D_i are not mutually independent, it is still possible

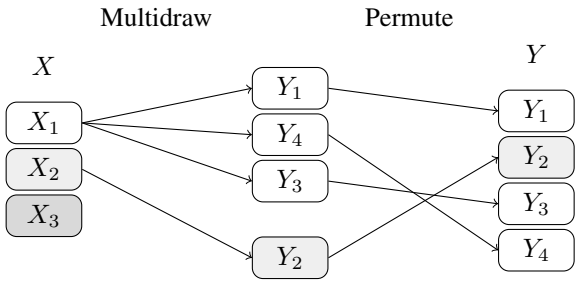

Fig. 3. Division of the noisy drawing channel into two sub-channels. The first part are M parallel multidraw channels with random number of draws D_i. The second part randomly permutes all sequences. The output sequences are indexed according to their appearance in the output Y.

to prove convergence of the drawing distribution. That is, $\frac{Q_d}{M} \to \mathsf{Poi}_c(d)$ converge jointly in probability. Therefore, the relative number of channels with d draws is asymptotically $\mathsf{Poi}_c(d)$. Since the capacity of the multidraw channel with d draws is C_d, it follows that the capacity of the first sub-channel is given by $\sum_d \mathsf{Poi}_c(d) C_d$. We now turn to discuss the influence of the second sub-channel. There are in total only $M - Q_0$ input sequences i which have been drawn at least once, i.e., $D_i > 0$. As the channel randomly permutes the sequences, the receiver has an uncertainty of roughly N^{M-Q_0} options to associate the output sequences with $M - Q_0$ drawn input sequences and $\binom{M}{M-Q_0}$ options to choose those positions with $D_i > 0$. A random coding argument then suggests that the rate loss associated with this uncertainty is roughly $((M - Q_0)\log N + O(M))/(ML) \to \beta(1 - \mathrm{e}^{-c})$. Note that the rigorous derivation of the capacity is more involved since a precise analysis of the effect of the permutation operation on the capacity is non-trivial. Details are provided in [16], [21].

Note that the above discussion suggests that the capacity of the channel obtained by replacing the BSC with an insertion/deletion channel in our model has, in a suitable parameter domain, a capacity equal to that in (3), where C_d is replaced by the capacity C_d^{ID} of a d-multidraw insertion/deletion channel. However this result remains unproven to date and is hindered due to the lack of a comprehensive understanding of even C_1^{ID}.

B. Discussion of the Capacity Expression

Having built an intuitive understanding of the origin of the capacity, we now proceed to discuss the capacity expression for different parameters regimes. Figure 4 shows the capacity for $\beta = 1/20$ and different values of 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. We observe that surprisingly already for $c = 5$, the capacity exhibits a very flat behavior, dropping only very slightly from its maximum at $p = 0$. This can be explained by the fact that in this case, the average number of times that a sequence is drawn is already large enough such that only a small rate loss is incurred by the errors in the sequences (c.f. Figure 2). It is observed that even for $p \to 0$, the capacity does not approach one. In fact, this result is known from [15], where it is shown that the capacity for this case is equal to $(1 - \beta)(1 - \mathrm{e}^{-c})$, which is a special case of Theorem 1. The reason for this behavior is that even in the noiseless case, a

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 09,2023 at 03:21:26 UTC from IEEE Xplore. Restrictions apply.

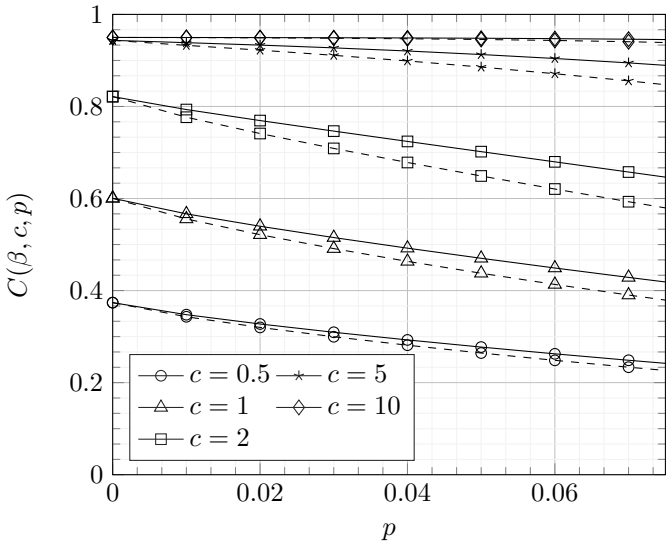

Fig. 4. Capacity of the noisy drawing channel for different values of c, over a range of error probabilities p, with $\beta = \frac{1}{20}$. The dashed lines show achievable rates for the case of suboptimal decoding with majority decisions.

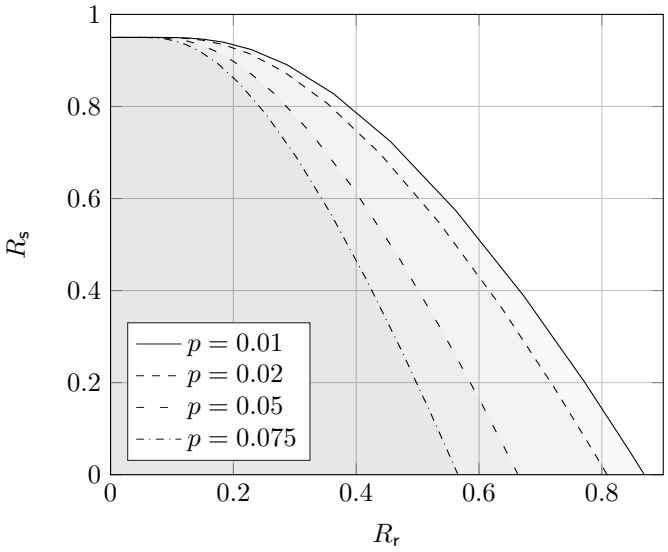

Fig. 5. Achievable storage and recovery rate region for different error probabilities p and $\beta = \frac{1}{20}$. The shaded regions highlight the achievable $(R_\mathsf{r}, R_\mathsf{s})$ pairs.

fraction of $1 - \mathrm{e}^{-c}$ input sequences are never observed at the receiver and it is further necessary to label the sequences with an index of length βL to combat the loss of ordering. On the other hand, if $c \to \infty$, the capacity approaches $1 - \beta$. This is intuitive, as in this case, each input sequence is drawn so many times, that the capacity of the multidraw channel is almost one for most sequences. Then, close to no error correction is required and it is sufficient to add an index of length βL to each sequence. Figure 4 also shows achievable rates using a suboptimal decoder based on majority voting. More precisely, for these curves we assume a decoder that, instead of using all output sequences to decode the original word, first performs a bitwise majority decision on all sequences that stem from the same input sequence. In the case where the majority is not unique, the decoder chooses the bit randomly. The maximum achievable rate is then measured for a channel with input X and output comprised of the results of the majority decision. The achievable rates in this case can be computed using (3) and replacing C_d with the capacity of a binary symmetric channel whose error probability is equal to the probability of a wrong decision after a majority vote on d bits. Clearly, due to a potential loss of information during the majority decision this decoding strategy is sub-optimal. However, we see that the overall rate loss with respect to the capacity is relatively small, depending on the channel parameters.

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 [15], [22]. In this regard, Figure 5 shows the regions of achievable $(R_\mathsf{r}, R_\mathsf{s})$ pairs for different error probabilities p and $\beta = \frac{1}{20}$. Notably, the region significantly flattens out for recovery rates R_r below approximately 0.1, which should be considered for efficient system design. We will elaborate on this in the next section.

IV. Design of Cost-Efficient DNA Archives

Having the capacity expression at hand, we will now present an optimization problem that will allow to design cost-efficient DNA storage systems. To start with, 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 that shall be stored, p is given by the synthesis and sequencing technologies. The costs associated with DNA-based data storage are mainly due 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

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

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 [15] we additionally incorporate the error probability into the system design. With this expression it is possible to optimize $\gamma(\beta, c, p)$ over c for given β and p. Currently the synthesis cost is a factor of roughly 10^4 larger than the sequencing cost [15] 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. Naturally it is possible to extend this cost optimization by incorporating, for example, that the costs are a function of the synthesis and sequencing quality p and then perform a joint optimization over c and p.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 09,2023 at 03:21:26 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, Sep. 2012. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1226355

[2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, Feb. 2013. [Online]. Available: http://www.nature.com/articles/nature11875

[3] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Sci. Rep.*, vol. 5, no. 1, p. 14138, Nov. 2015. [Online]. Available: http://www.nature.com/articles/srep14138

[4] S. M. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Sci. Rep.*, vol. 7, no. 1, p. 5011, Dec. 2017. [Online]. Available: http://www.nature.com/articles/s41598-017-05188-1

[5] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, C. N. Takahashi, S. Newman, H.-Y. Parker, C. Rashtchian, K. Stewart, G. Gupta, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze, and K. Strauss, "Random access in large-scale DNA data storage," *Nat. Biotechnol.*, vol. 36, no. 3, pp. 242–248, Mar. 2018. [Online]. Available: http://www.nature.com/articles/nbt.4079

[6] S. Chandak, J. Neu, K. Tatwawadi, J. Mardia, B. Lau, M. Kubit, R. Hulett, P. Griffin, M. Wootters, T. Weissman, and H. Ji, "Overcoming high nanopore basecaller error rates for DNA storage via basecaller-decoder integration and convolutional codes," in *Proc. Int. Conf. Acoust., Speech, Sig. Process.*, Barcelona, Spain, May 2020, pp. 8822–8826, iSSN: 2379-190X.

[7] R. G. Gallager, "Sequential decoding for binary channel with noise and synchronization errors," Arlington, VA, USA, Tech. Rep., Sep. 1961.

[8] M. C. Davey and D. J. C. Mackay, "Reliable communication over channels with insertions, deletions, and substitutions," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 687–698, Feb. 2001. [Online]. Available: http://ieeexplore.ieee.org/document/910582/

[9] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Exact reconstruction from insertions in synchronization codes," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2428–2445, Apr. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7809143/

[10] M. A. Sini and E. Yaakobi, "Reconstruction of sequences in DNA storage," in *Proc. Int. Symp. Inf. Theory*. Paris, France: IEEE, Jul. 2019, pp. 290–294. [Online]. Available: https://ieeexplore.ieee.org/document/8849740/

[11] M. Cheraghchi, R. Gabrys, O. Milenkovic, and J. Ribeiro, "Coded trace reconstruction," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6084–6103, Oct. 2020.

[12] A. Lenz, I. Maarouf, L. Welter, A. Wachter-Zeh, E. Rosnes, and A. Graell i Amat, "Concatenated codes for recovery from multiple reads of DNA sequences," in *Inf. Theory Workshop*, Riva del Garda, Italy, Apr. 2021, arXiv: 2010.15461. [Online]. Available: http://arxiv.org/abs/2010.15461

[13] S. R. Srinivasavaradhan, S. Gopi, H. D. Pfister, and S. Yekhanin, "Trellis BMA: Coded trace reconstruction on IDS channels for DNA storage," in *Proc. Int. Symp. Inf. Theory*, Melbourne, Australia, Jul. 2021.

[14] I. Shomorony and R. Heckel, "Capacity results for the noisy shuffling channel," in *Proc. Int. Symp. Inf. Theory*, Paris, France, Jul. 2019, pp. 762–766, arXiv: 1902.10832. [Online]. Available: http://arxiv.org/abs/1902.10832

[15] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental limits of DNA storage systems," in *Proc. Int. Symp. Inf. Theory*. Aachen: IEEE, Jun. 2017, pp. 3130–3134. [Online]. Available: http://ieeexplore.ieee.org/document/8007106/

[16] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Achieving the capacity of the DNA storage channel," in *Proc. Int. Conf. Acoust., Speech, Sig. Process.*, Barcelona, Spain, May 2020, pp. 8846–8850, iSSN: 2379-190X.

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

[18] M. Mitzenmacher, "On the theory and practice of data recovery with multiple versions," in *Proc. Int. Symp. Inf. Theory*. Seattle, WA: IEEE, Jul. 2006, pp. 982–986. [Online]. Available: http://ieeexplore.ieee.org/document/4036111/

[19] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Sci. Rep.*, vol. 9, no. 9663, Jul. 2019. [Online]. Available: http://www.nature.com/articles/s41598-019-45832-6

[20] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2331–2351, Apr. 2020.

[21] ——, "An upper bound on the capacity of the DNA storage channel," in *Proc. Inf. Theory Workshop*. Visby, Sweden: IEEE, Aug. 2019, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8989388/

[22] S. Chandak, K. Tatwawadi, B. Lau, J. Mardia, M. Kubit, J. Neu, P. Griffin, M. Wootters, T. Weissman, and H. Ji, "Improved read/write cost tradeoff in DNA-based data storage using LDPC codes," in *Proc. Annu. Allerton Conf. Commun. Control Comp.*, Monticello, IL, Sep. 2019, pp. 147–156.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 09,2023 at 03:21:26 UTC from IEEE Xplore. Restrictions apply.