Upper and Lower Bounds on the Capacity of the DNA-Based Storage Channel

Zihui Yan, Cong Liang, and Huaming Wu, *Senior Member, IEEE*

***Abstract*— The application of DNA as a powerful tool for storing digital information in chemically synthesized molecules has undergone continuous development. To explore its potential and limitations, we model the DNA storage channel as a cascade of a series of parallel and independent DNA noisy synchronization error channels and a shuffling-sampling channel, and derive novel lower and upper capacity bounds through a purely information-theoretic approach. Our results reveal the potential of DNA storage density and can be used to guide the design of error correction codes.**

***Index Terms*— DNA-based storage systems, synchronization error channels, channel capacity.**

I. Introduction

IN THE digital era of exploding quantities of data, breakthrough technologies are desired to achieve low-cost and low-consumption storage. DNA, the molecule encoding biological information, becomes an encouraging storage medium owing to its longevity and high information density. In recent years, researches and applications of this field have been widely concerned and studied [1], [2].

As shown in Fig. 1, in a typical DNA-based storage system, the data are first segmented into small pieces due to the limit of the synthesizing and sequencing technologies, then encoded into quaternary codewords via error correction codes [3], [4], and subsequently synthesized into DNA strands and stored. The data recovery process includes strand amplification, PCR experiments (polymerase chain reaction), random extraction, and sequencing, which finally output duplicate, disordered, and incorrect readouts.

Based on thorough assessments of the error sources and ratios under various experimental setups [5], there are typically three categories of errors in the aforementioned processes. Firstly, since the short DNA strands stored in test tubes are spatially disordered, the context of readouts cannot be intuitively derived from the order of receipt as in the communications field. Second, there are duplicates and unread DNA strands (called dropouts) due to the unstable and non-uniform number of synthesis and sequencing times of DNA strands. These two sources of error happen at the molecular level. Third, insertions, deletions, and substitutions are introduced at the nucleotide level during DNA synthesis and sequencing.

Manuscript received 18 July 2022; accepted 27 August 2022. Date of publication 29 August 2022; date of current version 11 November 2022. This work is supported by the National Key R&D Program of China (2020YFA0712102) and National Natural Science Foundation of China (12001401, 62071327). The associate editor coordinating the review of this letter and approving it for publication was M. Egan. *(Corresponding author: Huaming Wu.)*

The authors are with the Center for Applied Mathematics, Tianjin University, Tianjin 300072, China (e-mail: yanzh@tju.edu.cn; cong.liang@tju.edu.cn; whming@tju.edu.cn).

Digital Object Identifier 10.1109/LCOMM.2022.3202961

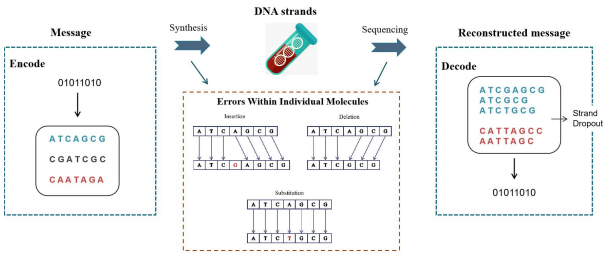

Fig. 1. The outline of the DNA-based storage system.

Nucleotide deletions are mainly caused by insufficient biochemical reactions. Whereas nucleotide insertions are mainly caused by overreaction, possibly a series of random nucleotides. These two errors are named synchronization errors in the classical communication channels. While nucleotide substitutions are mainly due to mutations. To characterize such errors, we model the transmission of each DNA strand as a DNA noisy synchronization error channel (***DNSEC***). Under different biochemical experimental conditions, the proportion of errors is various. Nevertheless, the common feature is a significant proportion of synchronization errors. The error analysis in [1] and [5] showed that synchronization errors account for at least half of the errors.

Unfortunately, synchronization errors have not been taken into consideration in previous studies of DNA storage channels; only substitutions have been studied. Shomorony and Heckel first characterized the random selecting and sequencing process as a shuffling-sampling channel (***SSC***) and derived the DNA storage channel capacity by cascading a binary symmetric channel (***BEC***) and a SSC [6]. Lenz *et al.* [7] extended this work and derived an upper bound for the cases when readout are duplicated and have substitution errors. However, these capacity bounds are overestimated significantly since synchronization errors are not accounted for.

For synchronization errors, the information coding theorem has been established by Dobrushin in [8]. See [9] for a recent survey. Previous work has deduced capacity bounds for many special synchronization error channels, such as the deletion channel (where $p_{\mathrm{i}} = p_{\mathrm{s}} = 0$) and the sticky channel (where $p_{\mathrm{d}} = p_{\mathrm{s}} = 0$ and insertions are the same as the input).

For deletion channels, Diggavi and Grossglauser obtained a lower bound of the deletion channel using an appropriate decoder, i.e., by detecting the unique correspondence of an output sequence and all subsequences of some original sequence [10]. For sticky channels, it is feasible to calculate its capacity by the equivalent capacity per unit cost [11], or only calculate its error-free capacity [12]. However, most previous work has focused on channels with a finite number of synchronization errors. Despite there being many methods

to estimate capacity bounds for these channels, there are no significant improvements in capacity bounds for channels with substitutions, deletions, and geometrically random insertions of which we are aware.

We clarify that the DNA channel is distinct from the noisy permutation channel introduced in [13]. Three assumptions of [13] that make its model different from ours: (i). no synchronization errors are introduced at the nucleotide level; (ii). no dropouts are introduced at the molecular level; (iii). the input alphabet is finite. Furthermore, our main problem is the fundamental limits of the asymptotic results in terms of achievable rate under a vanishing error probability formalism.

In this letter, we provide a new channel model for the DNA-based storage system and obtain its capacity. The channel capacity refers to the maximum of all rates where reliable transmission is possible, and it further represents the maximum number of bits that could be reliably stored in a single DNA molecule (called storage capacity). Our main contributions to this work are as follows.

- We model the DNA storage channel as a cascade of a set of DNA noisy synchronization error channels and a shuffling-sampling channel. Such a channel can characterize nucleotide errors, dropouts, and disorders, thereby it is more comprehensive and accurate than previous work. To the best of our knowledge, it is the first work on synchronization errors for DNA storage.
- We drive the upper and lower capacity bounds of the above channel through a purely information-theoretic approach. This work is the first to obtain a non-trivial capacity bound for DNSECs, and further obtains a novel capacity bound for DNA storage channels. While our work deals only with a few asymptotic results on information rates, we think that this model is useful for designing error correction codes for DNA storage in general.

II. DNA Storage Channel Model

A DNA strand is composed of four nucleotides (Adenine, Cytosine, Guanine, and Thymine) and can be treated as a sequence on a four-letter alphabet $\mathbf{\Sigma}$. We use upper case letters to represent random variables, while their realizations are depicted in lower case.

Based on the common characteristics of current biochemical technologies of DNA synthesis and sequencing, we conclude that the inputs of the DNA channel are quaternary codewords and the outputs are duplicate, disordered, and erroneous quaternary sequences. Since conventionally a clustering algorithm is applied to obtain consensus sequences before decoding [14], we assume that each original DNA strand corresponds to at most one received sequence, that is, no duplicate.

The mathematical model of the our channel is shown in Fig. 2. It can be seen as a cascaded channel, where the inner channel is the DNSEC, and the outer channel is the SSC. Let $X_1^n, X_2^n, \ldots, X_m^n$ denotes m inputs, where each $X_i^n \in \mathbf{\Sigma}^n$.

In the inner channel, the input X_i^n is translated into $Y_i^{N_i}$ via a DNSEC. Here, N denotes the number of received bits, which is a random variable depending on the realization of the insertion/deletion process. For illustration, the transition process characterizing a single use of the channel is shown in Fig. 3. Each input symbol is either deleted or

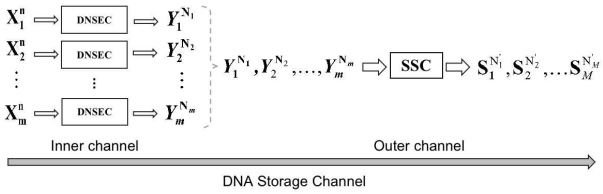

Fig. 2. The mathematical model of the DNA storage channel.

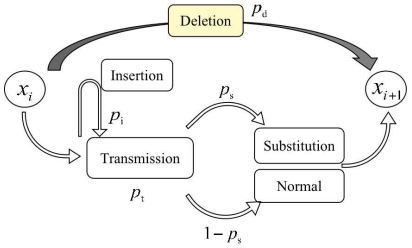

Fig. 3. A single use of the DNSEC.

transmitted. If transmitted, multiple random symbols might insert ahead of it with geometric probabilities, and the symbol may also be substituted. Here, since a deletion followed by an insertion makes a substitution error, we assume that the deletion does not occur after insertions in our model, which is different from models described in [15]. We denote the deletion, insertion, and substitution probabilities as p_{d}, p_{i}, and p_{s}, respectively, and assume that errors are independent and identically distributed (i.i.d.). The transmission probability is $p_{\mathrm{t}} = (1-p_{\mathrm{i}})(1-p_{\mathrm{d}})$ for normalization.

A single use of the DNSEC is characterized by an input alphabet $\mathbf{\Sigma}$, an output alphabet $\mathcal{Y} = \cup_{r=0}^{\infty}\mathbf{\Sigma}^r$, and a conditional probability distribution $p(\vec{y}|x)$ for every $x \in \mathbf{\Sigma}$ and $\vec{y} \in \mathcal{Y}$. The transition probabilities are

$$p(\vec{y}|x) = \begin{cases} p_{\mathrm{d}}, & \vec{y} = \epsilon; \\ (\frac{1}{4}p_{\mathrm{i}})^{r-1}p_{\mathrm{t}}(1-p_{\mathrm{s}}), & \vec{y} = *x \in \mathbf{\Sigma}^r, r \geq 1, \\ (\frac{1}{4}p_{\mathrm{i}})^{r-1}p_{\mathrm{t}}\frac{1}{3}p_{\mathrm{s}}, & \vec{y} = *x' \in \mathbf{\Sigma}^r, x' \neq x, r \geq 1, \end{cases} \quad (1)$$

where ϵ denotes the empty sequence and $*$ denotes a $(r-1)$-length sequence with random symbols. Given n inputs $x^n = x_1x_2\cdots x_n \in \mathbf{\Sigma}^n$, the output of each x_i is denoted as $\vec{y}_i$. The overall output is the in-order concatenation of $\vec{y}_i$ without delimiters, rewritten as $Y^N = (y_1, y_2, \ldots, y_N) \in \mathbf{\Sigma}^N$.

In the outer channel, $(Y_1^{N_1}, Y_2^{N_2}, \ldots, Y_m^{N_m})$ is the input, and the output is $(S_1^{N_1'}, S_2^{N_2'}, \ldots, S_M^{N_M'})$. The model of the SSC is an extension of the work by Shomorony and Heckel [6]. Specifically, it samples each input sequence independently with a uniform probability and outputs the samples in shuffled order. This results in that each output has no direct information to point to their corresponding original input, and some inputs are lost. Hence, the dropout error makes $M \leq m$, and each N_i' is independent of N_i due to the shuffle.

III. Capacity Bounds for DNA-Based Storage Channels

We use n and m to denote the length and number of original DNA strands, respectively. Then β is a positive constant which represents $\lim_{m,n\to\infty}\frac{n}{\log m}$. Let p_{d}, p_{i} and p_{s} denote

deletion, insertion, and substitution probabilities, respectively. Let q denote the probability that a given sequence is never sampled in the SSC. We use C_{DNA} to represent the DNA storage channel capacity. For notation convenience, let $(a)^+ \triangleq \max\{a, 0\}$, and $H(p) \triangleq -p\log p - (1-p)\log(1-p)$ with $0\log 0 = 0$.

Theorem 1: The capacity of the DNA storage channel can be bounded as

$$\left((1-q)\Big(\underline{C}_{inn} - \frac{1}{\beta}\Big)\right)^+ \leq C_{DNA} \leq \left((1-q)\Big(\overline{C}_{inn} - \frac{1}{\beta}\Big)\right)^+, \tag{2}$$

where

$$\begin{aligned}\underline{C}_{inn} &= (1-p_{\rm d})(2 - H(p_{\rm s}) - p_{\rm s}\log 3) - H(p_{\rm d}) \\ &\quad - \frac{1-p_{\rm d}}{1-p_{\rm i}} H(p_{\rm i}), \\ \overline{C}_{inn} &= (1-p_{\rm d})\left(2 - H(p_{\rm s}) - p_{\rm s}\log 3\right) - \frac{1-p_{\rm d}}{1-p_{\rm i}} H(p_{\rm i}).\end{aligned} \tag{3}$$

Proof: Since the transmission of DNA strands is modeled as the cascade channel introduced in Section II, we first calculate the capacity of the inner channel (i.e., the DNSEC), denoted as C_{inn}, and then extend the result to the cascade case.

According to Dobrushin [8], the capacity of synchronization error channels can be obtained by maximizing the mutual information. To be specific, since

$$\sum_{\vec{y}\in\mathcal{Y}} |\vec{y}| \cdot p(\vec{y}|x) = 0 \cdot p_{\rm d} + \sum_{r=1}^{\infty} r p_{\rm t} p_{\rm i}^{r-1} = \frac{1-p_{\rm d}}{1-p_{\rm i}}, \tag{4}$$

which shows that $\sum_{\vec{y}\in\mathcal{Y}} |\vec{y}| \cdot p(\vec{y}|x)$ is bounded as long as $p_{\rm i} \neq 1$. Thus, the capacity of the inner channel can be estimated via the mutual information, it follows that

$$C_{inn} = \lim_{n\to\infty} \max_{P_{X^n}} \frac{I(X^n; Y^N)}{n}. \tag{5}$$

We calculate the mutual information by revealing some side-information about the input to the receiver, drawing inspiration from works on [16]. Firstly, through estimating the mutual information when inputs are independent and uniformly distributed (i.u.d.), we obtain the lower/upper bound of C_{inn}, denoted as $\underline{C}_{inn}$.

Lemma 1: The lower bound of the DNSEC is

$$\underline{C}_{inn} = \left((1-p_{\rm d})(2 - H(p_{\rm s}) - p_{\rm s}\log 3) - H(p_{\rm d}) - \frac{1-p_{\rm d}}{1-p_{\rm i}} H(p_{\rm i})\right)^+,$$

and can be achieved with an i.u.d. input.

Proof: The proof follows the assumption that the input distribution is uniform. We first introduce an auxiliary sequence $D^n = (D_1, D_2, \ldots, D_n)$, where $D_i \in \mathbb{Z}$ uniquely determines the length of $\vec{Y}_i$. This auxiliary sequence is not observed for the decoder. And $D_1, D_2, \ldots, D_n$ are i.i.d. with the probability distribution

$$Pr[D = r] = \begin{cases} p_{\rm d}, & r = 0; \\ p_{\rm t} p_{\rm i}^{r-1}, & r \geq 1. \end{cases} \tag{6}$$

According to the chain rule of information, we have

$$I(X^n; Y^N) = I(X^n; Y^N, D^n) - I(X^n; D^n|Y^N). \tag{7}$$

Here, since D^n indicates deletion and insertion error positions, $X^n \to (Y^N, D^n)$ induces a memoryless channel with the erasure probability $p_{\rm d}$ and the substitution probability $p_{\rm s}$, it follows that

$$I(X^n; Y^N, D^n) = n(1-p_{\rm d})(2 - H(p_{\rm s}) - p_{\rm s}\log 3), \tag{8}$$

where the equal sign is met since X^n and D^n are independent, and $X_1, X_2, \ldots, X_n$ are i.u.d..

For the second term of (7), we have

$$I(X^n; D^n|Y^N) = H(D^n|Y^N) - H(D^n|X^n, Y^N). \tag{9}$$

With an i.u.d. input, the output is also i.u.d., thereby the only information obtained from Y^N about D^n is the length of the overall output, which is equivalent to the sum of D^n. Hence,

$$H(D^n|Y^N) = \sum_{j=0}^{\infty} Pr[N = j] H(D^n|N = j). \tag{10}$$

To obtain $H(D^n|N = j)$, we consider the method of types. For any sequence $d^n = (d_1, d_2, \ldots, d_n)$, which satisfies $\sum_{i=1}^n d_i = j$, denote its type as $P_{d^n}^{(j)} = (P_{d^n}^{(j)}(0), P_{d^n}^{(j)}(1), \ldots, P_{d^n}^{(j)}(j))$, where $P_{d^n}^{(j)}(i) = N(i|d^n)/n$ (i.e., $N(i|d^n)$ is the number of times the symbol i occurs in the sequence d^n). Let $\mathcal{P}_n^{(j)} = \left\{\left(P_n^{(j)}(0), P_n^{(j)}(1), \ldots, P_n^{(j)}(j)\right) \in \mathcal{R}^{j+1} : P_n^{(j)}(i) \geq 0, \sum_{i=1}^j P_n^{(j)}(i) = 1, \sum_{i=1}^j i P_n^{(j)}(i) = j/n\right\}$ denote the set of types with denominator n, which is the subset of the probability simplex in $\mathcal{R}^{j+1}$. It is obvious that $P_{d^n}^{(j)} \in \mathcal{P}_n^{(j)}$. Let $\triangle(P) = \{x^n \in \mathbb{N}^n : P_{x^n} = P\}$ denote the type class of P, it follows that $\left|\cup_{d^n:\sum d_i = j} \triangle(P_{d^n}^{(j)})\right| = \left|\mathcal{P}_n^{(j)}\right| = \binom{n+j-1}{n-1}$. We now use the size and the probability of type classes to evaluate (10),

$$(10) = \sum_{j=0}^{\infty} \sum_{P_{d^n}^{(j)} \in \mathcal{P}_n^{(j)}} Pr[P_{d^n}^{(j)}] \log\left|\mathcal{P}_n^{(j)}\right|. \tag{11}$$

For any type P_{d^n}, the probability of the type class $\triangle(P_{d^n})$ is $2^{-nD(P_{d^n}||P_D)}$. According to the law of large numbers, we have $D(P_{d^n}||P_D) \to 0$ with probability 1. It follows that the probability of the strongly typical set $A_\epsilon^{(n)} = \left\{d^n : \left|N(i|d^n)/n - p_{\rm d}(i)\right| < \epsilon\right\}$ goes to 1 as $n \to \infty$. Thus, we can use P_D in (6) to estimate the properties of the sequence D^n, it follows that

$$\begin{aligned}(11) &= \log\left((1-\epsilon)2^{n(H(D)-\epsilon)}\right) + o(n) \\ &= nH(D) + o(n) \\ &= nH(p_{\rm d}) + \frac{1-p_{\rm d}}{1-p_{\rm i}} H(p_{\rm i}) + o(n).\end{aligned} \tag{12}$$

According to the non-negativity of entropy, we have $H(D^n|X^n, Y^N) \geq 0$ to estimate the upper bound of (9).

To sum up, the capacity can be lower bounded by plugging the results of (8) and (9) into (7), and is achievable via an i.u.d. input. □

Next we drive the upper bound of the inner channel, denoted as $\overline{C}_{inn}$.

Lemma 2: The upper bound of the DNSEC is

$$\overline{C}_{inn}=\left((1-p_{\rm d})\,(2-H(p_{\rm s})-p_{\rm s}\log 3)-\frac{1-p_{\rm d}}{1-p_{\rm i}}H(p_{\rm i})\right)^{+}.$$

Proof: As shown in [16], the side-information D^n will not decrease the capacity due to $X^n \to (\vec{Y}_1, \vec{Y}_2, \ldots, \vec{Y}_n) \to Y^N$ forms a Markov chain. Thus the mutual information can be calculated via

$$\begin{aligned} I(X^n;Y^N) &\le I(X^n;\vec{Y}_1,\vec{Y}_2,\ldots,\vec{Y}_n) \\ &= \sum_{i=1}^{n}\Big(H(\vec{Y}_i)-H(\vec{Y}_i|X_i)\Big). \end{aligned} \tag{13}$$

Since $\vec{Y}_i$ has a probability $p_{\rm d}$ to be an erasure, it follows that

$$H(\vec{Y}_i)\le 2\frac{1-p_{\rm d}}{1-p_{\rm i}}+H(p_{\rm d}). \tag{14}$$

Then, according to the transition probability (1), we have

$$\begin{aligned} &H(\vec{Y}_i|X_i) \\ &\quad= -p_{\rm d}\log p_{\rm d}-\sum_{r=0}^{\infty}\Big(p_{\rm i}^r p_{\rm t}(1-p_{\rm s})\log(\frac{p_{\rm i}}{4})^r p_{\rm t}(1-p_{\rm s}) \\ &\qquad+p_{\rm i}^r p_{\rm t}p_{\rm s}\log(\frac{p_{\rm i}}{4})^r p_{\rm t}\frac{p_{\rm s}}{3}\Big) \\ &\quad= H(p_{\rm d})+\frac{1-p_{\rm d}}{1-p_{\rm i}}\,(H(p_{\rm i})+2p_{\rm i}) \\ &\qquad+(1-p_{\rm d})\,(H(p_{\rm s})+p_{\rm s}\log 3)\,. \end{aligned}$$

Hence, it follows that,

$$\begin{aligned} &\lim_{n\to\infty}\frac{I(X^n;Y^N)}{n} \\ &\quad\le (1-p_{\rm d})\,(2-H(p_{\rm s})-p_{\rm s}\log 3)-\frac{1-p_{\rm d}}{1-p_{\rm i}}H(p_{\rm i}). \end{aligned} \tag{15}$$

□

Let us now address the evaluation of the above capacity bounds. When $p_{\rm i}=0$, Lemma 1 coincides with the lower bound provided in [10]. However, the decoding technique provided in [10] (i.e., a common subsequence detection rule) cannot be applied to our model due to random insertions. An interesting finding is that when $p_{\rm d}=p_{\rm s}=0$, our proof shows that the i.u.d. input can achieve the capacity of geometric random-insertion channels. The study of this channel capacity is lacking. Our proof is reasonable because the uniform distribution maximizes the entropy. And D^n can be determined through trellis-structure decoding [17], so that $H(D^n|X^n,Y^N)$ tends to zero with the decoding error probability goes to zero.

Armed with the above descriptions of the inner channel, we now drive the overall channel capacity. In the context of concatenated error correction coding schemes, the outer decoding corrects the residual of the inner decoding. Based on this consideration, we use R^e_{inn} to represent the rate of the inner code, in which the average error probability is denoted as P^e_{inn}. We have

$$nR^e_{inn}\le 1+P^e_{inn}nR^e_{inn}+I(X^n;Y^N). \tag{16}$$

Back to the overall channel, we use $X^{mn}=X_1^nX_2^n\ldots X_m^n$ to represent the input, and $S^{MN}=S_1^{N_1}S_2^{N_2}\ldots S_M^{N_M}$ to represent the output. Let R_{all} denote the achievable rate of the overall channel with the overall average error probability P^e_{all} goes to zero. Given that S^{MN} is output out-of-order, we introduce a side-information $\Pi^M=(\Pi_1,\Pi_2,\ldots,\Pi_M)$ to represent the index of S^{MN} (i.e., $S_i^{N_i(n)}$ is transmitted from $X^n_{\Pi_i}$). Hence,

$$\begin{aligned} mnR_{all} &\overset{(a)}{=} H(X^{mn}) \\ &\overset{(b)}{\le} I(X^{mn};S^{MN})+H(X^{mn}|\hat{X}^{mn}) \\ &\overset{(c)}{=} I(X^{mn},\Pi^M;S^{MN})-I(\Pi^M;S^{MN}|X^{mn}) \\ &\quad+H(X^{mn}|\hat{X}^{mn}), \end{aligned} \tag{17}$$

where (a) follows from the assumption that X^{mn} is uniform over $\{1,2,\ldots,2^{mnR}\}$, (b) is the data-processing inequality, and (c) is the chain rule for information.

We now calculate the first term of (17). Given the index sequence Π^M, $(X^{mn},\Pi^M)\to S^{MN}$ can be seen as the DMC with the input alphabet $\mathbf{GF}(2^{nR^e_{inn}})$ and the substitution probability P^e_{inn}. For this channel, it follows that

$$\begin{aligned} &I(X^{mn},\Pi^M;S^{MN}) \\ &\le (1-q)m\left(nR^e_{inn}-H(P^e_{inn})-P^e_{inn}\log(2^{nR^e_{inn}}-1)\right), \end{aligned} \tag{18}$$

where the equal sign is met when input are i.u.d.. It could be achievable when symbols in each X^n are i.u.d..

For the second term of (17), we have $I(\Pi^M;S^{MN}|X^{mn})=H(\Pi^M|X^{mn})-H(\Pi^M|S^{MN},X^{mn})$. Here,

$$\begin{aligned} &H(\Pi^M|X^{mn}) \\ &\overset{(a)}{=}\sum_{i=1}^{m}Pr[M=i]\log\frac{m!}{(m-i)!} \\ &\overset{(b)}{=}\sum_{i=1}^{m}Pr[M=i]\left(i\log m+(m-i)\log\frac{m}{m-i}\right)+o(m) \\ &\overset{(c)}{=}(1-q)m\log m+o(m), \end{aligned} \tag{19}$$

where (a) follows the fact that Π^M is independent of X^{mn} and Π^M is chosen uniformly at random from all vectors in $\{1,\ldots,m\}$ with distinct elements, (b) follows from the Stirling approximation, and (c) is Jensen's inequality,

$$\sum_{i=1}^{m}Pr[M=i](m-i)\log\frac{m}{m-i}\le(1-q)m\log\frac{1}{q}=o(m).$$

Our last task is to estimate $H(\Pi^M|S^{MN},X^{mn})$. As the idea provided in [5], given X^{mn} and S^{MN}, we estimate the permutation $\hat{\Pi}^M$ by mapping each output and corresponding input. To be specific, it is assumed that $\Pi_i=j$ if the decoding result of S_i^N is X_j^n. From S^{MN}, we make an estimate $\hat{X}^{mn}$, and let P^e_{all} be the maximum error probability of the overall decoding, so that $Pr[X^{MN}\ne\hat{X}^{MN}]=P^e_{all}$. Under this error probability, we have $Pr[\Pi^M\ne\hat{\Pi}^M]=P^e_{all}$. From Fano's inequality, it follows that

$$\begin{aligned} H(\Pi^M|S^{MN},X^{mn}) &\le H(\Pi^M|\hat{\Pi}^M) \\ &= H(P^e_{all})+P^e_{all}(1-q)m\log m. \end{aligned} \tag{20}$$

For the last term of (17), according to Fano's inequality, we have

$$H(X^{mn}|\hat{X}^{mn})\le 1+mnR_{all}P^e_{all}. \tag{21}$$

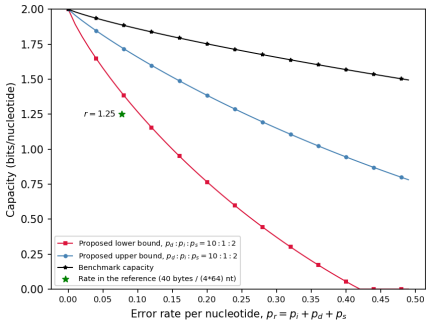

Fig. 4. Capacity bounds, when $q = 1/\beta = 0$. The star mark points to the storage capacity in [1].

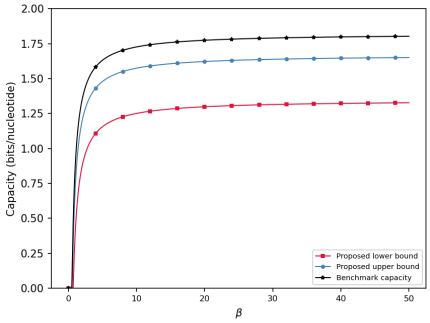

Fig. 5. Capacity bounds, when $q = 4.5\%$, $p_{\mathrm{d}} = 6.3\%, p_{\mathrm{i}} = 0.45\%$, and $p_{\mathrm{s}} = 0.94\%$.

Finally, by plugging results of (16), (18), (19), (20), and (21) into (17), the achievable rate of the DNA storage channel (2) can be obtained as $m, n \to \infty$ and $P_{all}^{e} \to 0$. □

IV. Numerical Results

In this section, comparisons between the proposed bounds and the existing DNA storage channel bounds are given. The error probabilities we refer to come from two recent instructive letters. Nguyen *et al.* [1] stored DNA data by nanoscale electrode wells, in which error analysis showed that $p_{\mathrm{d}} = 6.3\%, p_{\mathrm{i}} = 0.45\%$, and $p_{\mathrm{s}} = 0.94\%$. Winston *et al.* [2] presented a new combinatorial PCR method, in which a filtering process resulted in an average of 4.5% of strands being lost.

We first focus on channels with $q = 0$ and $1/\beta = 0$, that is, no information loss in the outer channel. And we assume $p_{\mathrm{d}} : p_{\mathrm{i}} : p_{\mathrm{s}} = 10 : 1 : 2$ to simulate the error probabilities. The capacity bounds (2) are reported in Fig. 4. The proposed bounds significantly tighten the reference benchmark, namely, the capacity bound of the DNA storage channel from [6]. These curves indicate that DNA channel capacity is appreciably overestimated if synchronization errors are neglected.

Next, we focus on capacity loss due to the disordered permutation in the outer channel. Consistent with the conclusion in [6] and [7], our proof shows that a simple index-based coding scheme is optimal for the outer code even taking synchronization errors. There has also been research on how to design these indexes, namely, primer sets [18]. The index leads to a drop in capacity by at least $1/\beta$. For large β, information is difficult to transmit reliably over the channel, as shown in Fig. 5. Thus, the capacity in (2) is only non-trivial if the sequence length scales as $n = \Theta(\log m)$.

V. Conclusion

In this work, we presented a new model for DNA storage channel, which is a cascade of a series of parallel and independent channels and a shuffling-sampling channel, and derived its lower and upper capacity bounds. The presented upper bound was obtained by exploiting an auxiliary system where suitable side information is revealed to the receiver, and by computing the relevant capacity via suitable information-theoretical inequalities. The lower bound was obtained by exploiting the same auxiliary system and was achievable via an i.u.d. input. To the best of our knowledge, it is the first work on the DNA storage channel that considers synchronization errors, which provided tighter capacity bounds on DNA storage channels. It further facilitates the exploration of fundamental theoretical questions, e.g., establishing error correction coding schemes for the DNA-based storage system.

References

[1] B. H. Nguyen *et al.*, "Scaling DNA data storage with nanoscale electrode wells," *Sci. Adv.*, vol. 7, no. 48, Nov. 2021, Art. no. eabi6714.
[2] C. Winston, L. Organick, D. Ward, L. Ceze, K. Strauss, and Y.-J. Chen, "A combinatorial PCR method for efficient, selective oligo retrieval from complex oligo pools," *ACS Synth. Biol.*, vol. 11, no. 5, pp. 1727–1734, Feb. 2022.
[3] W. Song, K. Cai, and K. A. S. Immink, "Sequence-subset distance and coding for error control in DNA-based data storage," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6048–6065, Jun. 2020.
[4] P. Mishra, C. Bhaya, A. K. Pal, and A. K. Singh, "Compressed DNA coding using minimum variance Huffman tree," *IEEE Commun. Lett.*, vol. 24, no. 8, pp. 1602–1606, Aug. 2020.
[5] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Sci. Rep.*, vol. 9, no. 1, p. 9663, Dec. 2019.
[6] I. Shomorony and R. Heckel, "DNA-based storage: Models and fundamental limits," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 3675–3689, Jun. 2021.
[7] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "An upper bound on the capacity of the DNA storage channel," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Aug. 2019, pp. 1–5.
[8] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," *Problemy Peredachi Informatsii*, vol. 3, no. 4, pp. 18–36, 1967.
[9] M. Cheraghchi and J. Ribeiro, "An overview of capacity results for synchronization channels," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 3207–3232, Jun. 2021.
[10] S. N. Diggavi and M. Grossglauser, "Information transmission over a finite buffer channel," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2000, p. 52.
[11] M. Mitzenmacher, "Capacity bounds for sticky channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 1, pp. 72–77, Jan. 2008.
[12] M. Kovacevic, "Zero-error capacity of duplication channels," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6735–6742, Oct. 2019.
[13] A. Makur, "Coding theorems for noisy permutation channels," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 6723–6748, Nov. 2020.
[14] C. Rashtchian *et al.*, "Clustering billions of reads for DNA data storage," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3362–3373.
[15] 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.
[16] H. Mercier, V. Tarokh, and F. Labeau, "Bounds on the capacity of discrete memoryless channels corrupted by synchronization and substitution errors," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4306–4330, Jul. 2012.
[17] J. Hu, T. M. Duman, and M. F. Erden, "On the information rates of channels with insertion/deletion/substitution errors," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 956–960.
[18] Y. M. Chee, H. M. Kiah, and H. Wei, "Efficient and explicit balanced primer codes," *IEEE Trans. Inf. Theory*, vol. 66, no. 9, pp. 5344–5357, Sep. 2020.