On Models and Capacity Bounds for DNA-based Storage Channels

Zihui Yan and Cong Liang

Abstract

In recent years, DNA-based systems have become a promising medium for long-term data storage. There are two layers of errors in DNA-based storage systems. The first is the dropouts of the DNA strands, which has been characterized in the shuffling-sampling channel. The second is insertions, deletions, and substitutions of nucleotides in individual DNA molecules. In this paper, we describe a DNA noisy synchronization error channel to characterize the errors in individual DNA molecules. We derive non-trivial lower and upper capacity bounds of the DNA noisy synchronization error channel based on information theory. By cascading these two channels, we provide theoretical capacity limits of the DNA storage system. These results reaffirm that DNA is a reliable storage medium with high storage density potential.

Index Terms

DNA storage, Synchronization error channels, Substitution errors, Capacity bounds.

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, many research and application of this field have been widely concerned and studied [1], [2], [3], [4], [5]. In a typical DNA-based storage system, the data are fragmented into small pieces and stored in short DNA segments due to the limit of the synthesizing and sequencing length. There are three main procedures in the DNA-based data storage and extraction process. The first is to synthesize the encoded message symbols into DNA molecules, the second is to conserve the synthesized DNA strands, and the

last is to sequence the DNA strands to recover the stored message sequences (Fig.1). There are two main error sources in the DNA storage system [6]. Firstly, DNA strands stored in a test tube are spatially disordered. In the data extraction process, DNA strands were randomly selected and sequenced, which means that there are chances that some strands were not sequenced. We term them strand-dropout errors, and a channel model with such errors has been described as a shuffling-sampling channel (*SSC*) [7]. Secondly, the sequencing and synthesis procedures may also bring insertions, deletions, and substitutions of nucleotides in individual DNA molecules. The insertions and deletions are synchronization errors, which were called indels collectively [8], [9]. Whereas the substitutions in DNA molecules corresponds to the additive errors in the classical communication channels. We call a channel with such error types the DNA noisy synchronization error channel (*DNSEC*). Thus, the DNA storage channel could be described as a cascaded channel of an SSC and a DNSEC. As a result, most of the DNA-based storage systems utilized a concatenated coding scheme: an outer code designed to recover the message on dropout strands, and an inner code to correct the errors within individual DNA molecules.

Fig. 1. A typical DNA storage system. Binary messages are encoded and synthesized into DNA strands. DNA strands need to be sequenced and decoded to recover the original message. In this process, some strands may not be sequenced. Also, the DNA synthesis and sequencing procedures may bring additive errors and synchronization errors in individual molecules.

Although the feasibility of storing information in synthesized DNA molecules have been investigated extensively, the capacity of the DNA storage channel remains an open question. As in previous studies, channel capacity refers to the maximum of all rates where reliable transmission is possible, and it can represent the maximum number of bits reliably stored in a single DNA molecule [7]. The main study

of this paper is to provide characterizations of the DNA storage channel. We consider the asymptotic regime with vanishing decoding error probability as the block size and sequence length increase. Note that we will not go into detail on how to construct error correction codes for DNA-based storage channels. Shannon's noise channel coding theory shows that we can always find a coding structure to achieve reliable transmission when the code rate is less than the channel capacity. On the contrary, when the code rate exceeds the capacity, the decoding error probability will converge to 1 as the block length increases.

This paper studies the DNSEC with the input alphabet $\mathcal{K} = \{0, 1, \dots, K - 1\}$ for DNA-based storage systems. We present a lower and an upper bound on the channel capacity based on information theory. In particular, when limiting $K = 4$, it is the DNA storage with four nucleotides (Adenine, Cytosine, Guanine, and Thymine). This paper is organized as follows. In Section II, we describe channel models with deletions and insertions termed DNA synchronization error channel (*DSEC*), then bring in substitutions to describe the channel model of DNSEC. In Section III, we derive the lower and the upper capacity bound of the DSEC. And in Section IV, we derive the lower and the upper capacity bounds of the DNSEC by cascading the DSEC and the discrete memoryless channel (*DMC*). Then we combine the results with the SSC to construct a complete DNA storage channel and obtain its capacity bound in Section V. Finally, we discuss the capacity limits of the DNA-based storage systems and show that the DNA-based storage system could be a reliable and high-density storage medium. We discuss some open issues, such as tighter capacity bounds and error correction code constructions in Section VI.

A. Previous Works

Studies have investigated the capacity of the SSC as well as some modifications to this channel. Shomorony and Heckel [10] first described the SSC that characterizes the random sampling process in the DNA-based storage system and derived its capacity, assuming that each sampled readout is error-free. Later [7], they extend the results to consider substitution errors in the readouts and derived the capacity of the DNA storage channel while assuming that the probability of DNA strands being sampled is a simple Bernoulli distribution. In practice, the DNA synthesizing and amplification process generates multiple copies of each DNA sequence. Each synthesized DNA strand could be sequenced multiple times with possibly different errors. As a result, [11], [12], derived an upper bound for the cases when sequences are drawn multiple times with the presence of substitution errors, and proposed a decoding method that could achieve that bound. It remains an open question how the capacity bounds might change when deletions and insertions within the individual DNA molecules are also considered.

Efforts have been made to derive the capacity bounds of the binary synchronization channels as well. In this paper, synchronization channels (**SC**) or synchronization error channels (**SEC**) refer to a class of channels that contain synchronization errors such as insertions and deletions [8], [9]. We also call a channel containing both synchronization errors and additive errors a noisy synchronization error channel (**NSEC**) [13]. Gallager [14] first obtained a lower bound of SECs that allow deletion and single insertion in each position of bits. Then lower bounds have been proposed for other cases, such as deletion channels, sticky channels, and deletion-duplication channels. For example, Diggavi and Grossglauser studied the lower bound of SECs containing deletions using an appropriate decoder. Given the output $\vec{y}$, the decoder checks whether $\vec{y}$ is a subsequence of only one codeword. If so, output this codeword, but if not, the decoder simply declares an error [13]. They found that the order-one Markov input could achieve a higher rate than the independent and uniform distributed input. This lower bound was improved by Drinea and Mitzenmacher [15], [16]. On the other hand, Ullman derived the zero-error capacity upper bound for deletion channels [17]. And the first non-trivial capacity upper bound for deletion channels was obtained by Diggavi [18], which assumes some extra side information about the input at the receiver (what is called genie-aided information). However, there is still a lack of researches on the synchronization error channel models as DNSEC.

II. THE DNA NOISY SYNCHRONIZATION CHANNEL MODELS

In this section, we construct the mathematical model for the DNSEC. In practice, the nucleotides are synthesized one by one by a chemical process in currently utilized synthesis strategies, such as light via optically sensitive protection groups [19], and electrodes via pH-sensitive protection groups [20]. This indicates that we could assume that nucleotides at different spots are independent to each other. Besides, the error distribution of DNA sequencing depends on the used techniques. For example, the most frequently utilized sequencing platforms are the Illumina and the nanopore sequencing [4], [21]. According to error statistics, DNA synthesis leads to deletions and a few insertions, and DNA sequencing leads to substitutions and very few insertions and deletions. Besides, long-term storage brings a few substitutions. And the error analysis of experiment by Organick et al. showed that the average error rate per position is almost 0.6%, with 0.4% substitutions, 0.2% deletions and 0.04% insertions [4]. We assume that the errors are independent and identically distributed (i.i.d.) at different positions of the DNA strands.

Indeed, the SEC is not memoryless, which means that there are correlations between different channel output symbols even if the input symbols are i.i.d.. They can not be corrected by techniques designed

for memoryless channels, such as Binary Symmetric Channel (BSC) and Binary Erasure Channel (BEC). Moreover, most of the theorems for memoryless channels based on Shannon Information Theory, are failed when applied to the SEC, which brings a great challenge to analyze this channel model. As a result, we first discuss the channel containing only synchronization errors then cascade it with the channel taking additive errors to construct the complete DNSEC model.

We study the following DNA noisy synchronization channel models:

- A DNA synchronization error channel (**DSEC**) is a simplified DNA storage inner channel with only deletions and insertions. It is characterized by an input alphabet $\mathcal{K} = \{0, 1, \dots, K-1\}$, an output alphabet $\mathcal{Y} = \cup_{l=0}^{\infty} \{0, 1, \dots, K-1\}^l$, and a conditional probability distribution $p(\vec{y}|x)$ for every $x \in \mathcal{K}$ and $\vec{y} \in \mathcal{Y}$. Here $\vec{y}$ is a vector with undefined length. For n input symbols $X^n = (x_1, x_2, \dots, x_n) \in \mathcal{K}^n$, the channel outputs $\vec{Y} = (\vec{y}_1, \vec{y}_2, \dots, \vec{y}_n)$, and $p(\vec{Y}|X^n) = \prod_{i=1}^n p(\vec{y}_i|x_i)$. Note that since the length of $\vec{y}_i$ varies, it may not be possible to segregate each $\vec{y}_i$ from $\vec{Y}$. Hence, the output is the concatenation of all $\vec{y}_i$, which we rewrite as $Y^{N(n)} = (y_1, y_2, \dots, y_{N(n)})$. We denote the deletion probability p_d , the insertion probability p_i , and the transmission probability $p_t = (1 - p_i)(1 - p_d)$ for normalization. Then the transition probabilities

$$p(\vec{y}|x) = \begin{cases} p_d, & \vec{y} = \epsilon; \\ p_t p_i^{r-1}, & \vec{y} = *x \in \mathcal{K}^r, r \geq 1, \end{cases} \quad (1)$$

where ϵ denotes the empty sequence and $*$ denotes a $(r-1)$ -length sequence with random symbols. In this channel, a symbol is either deleted or transmitted. If the symbol is transmitted, multiple insertions with geometric probabilities might happen ahead of it.

- A DNA noisy synchronization error channel (**DNSEC**) is a channel that contains synchronization errors and additive errors. If a symbol is transmitted, substitution may occur with probability p_s . We assume that the substitution probabilities of input symbols are also i.i.d.. Thus, the DNSEC is a cascaded model of the DSEC and a DMC with substitution. The transmission process is illustrated in Fig. 2.

Fig. 2. The transmission probabilities of a single symbol in the DNSEC. "Deletion" is to delete x_i and output an empty sequence. "Insertion" is to output one random symbol. "Transmission" is to output x_i with a fraction p_s for substitution. In brief, the errors occurred to x_i can be represented as a single color path from x_i to x_{i+1} , which we call the transmission track of symbol x_i .

The notations we used throughout the paper were summarized in table I.

III. CAPACITY BOUNDS OF THE DSEC

We now focus on analyzing the capacity of the DSEC. In his seminal work, Shannon [22] proved that the capacity of the DMC can be characterized by the maximal mutual information of the input and received message under all input distributions. Dobrushin showed that an analogous theorem also holds for synchronization channels [23].

Theorem 1. [Dobrushin] *Consider a synchronization error channel such that some real constants $0 < c_1 \leq c_2$ exists such that*

$$c_1 \leq \sum_{\vec{y} \in \mathcal{Y}} |\vec{y}| \cdot p(\vec{y}|x) \leq c_2, \quad (2)$$

for all $x \in \mathcal{K}$. Then, the capacity C of this channel could be calculated by

$$C = \lim_{n \rightarrow \infty} \max_{P_{X^n}} \frac{I(X^n; Y^{N(n)})}{n}, \quad (3)$$

where the maximum should be taken over all distributions P_{X^n} , and $Y^{N(n)}$ denotes the corresponding output sequence of X^n .

Consider the DSEC model, we have

$$\sum_{\vec{y} \in \mathcal{Y}} |\vec{y}| \cdot p(\vec{y}|x) = 0 \cdot p_d + \sum_{r=1}^{\infty} r p_t p_i^{r-1} = \frac{1 - p_d}{1 - p_i}. \quad (4)$$

TABLE I
NOTATION SUMMARY

Notations	Description
K	Alphabet size of the input.
p_d	Probability of deletion.
p_i	Probability of single insertion.
p_t	Probability of transmission ($p_t = (1 - p_d)(1 - p_i)$).
p_s	Probability of substitution.
q	Probability that a given strand is never sampled.
n	Input sequence length.
m	Quantity of DNA strands.
i	Sample index of symbol.
$N(n)$	Output sequence length.
X^n	Input sequence with length- n .
$Y^{N(n)}$	Output sequence of the DSEC with length- $N(n)$.
$\vec{Y}_i$	Output sequence of X_i .
D_i	Length of $\vec{Y}_i$ of the DSEC.
D^n	Auxiliary sequence that reveals the transmission track of the DSEC.
$N_{D^n}(i)$	Index of the last output symbol of X_i under transmission track D^n .
I	Index set of transmitted symbols of X^n in $Y^{N(n)}$ ($\mathbb{E}(I) = (1 - p_d)n$).
$Y^{N(n)} _I$	Transmitted symbol sequence.
$Z^{N(n)}$	Output sequence of the DNSEC with length- $N(n)$.
E_i	Additive error type of i^{th} symbol in $Y^{N(n)} _I$ of the DMC ($Pr[E_i \neq 0] = p_s$).
X^{mn}	Input of the DNA storage channel $X^{mn} = (X_1^n, X_2^n, \dots, X_m^n)$.
Z^{MN}	Output of the DNA storage channel $Z^{MN} = (Z_1^N, Z_2^N, \dots, Z_M^N)$.
S_i	Real correspondence between Z^{MN} and X^{mn} ($Z_i^N = X_{S_i}^n$).
g	Decoding function.
$\hat{S}^n$	Estimated index function of the SSC.
$P_e^{(m)}$	Decoding error probability of the DNA storage channel.
$ x $	Length of a sequence x .
$(x)^+$	$(x)^+ = \max\{x, 0\}$.
$H(p)$	Binary entropy function $H(p) = -p \log p - (1 - p) \log(1 - p)$.

Thus $\sum_{\vec{y} \in \mathcal{Y}} |\vec{y}| \cdot p(\vec{y}|x)$ is bounded as long as $p_d, p_i \neq 1$. The condition in Equation (2) holds. However, it is difficult to characterize the mutual information in Equation (3) to determine the exact capacity of DSEC.

To find the capacity bounds of DSEC, we introduce an auxiliary sequence $D^n = (D_1, D_2, \dots, D_n)$, where $D_i \in \mathbb{Z}$ is the length of $\vec{Y}_i$, which uniquely determines the transmission track of X_i . Note that the auxiliary sequence is not observed in the DSEC. And $(D_1, D_2, \dots, D_n)$ are i.i.d. with probability distribution

$$Pr[D = r] = \begin{cases} p_d, & r = 0; \\ p_t p_i^{r-1}, & r \geq 1. \end{cases} \quad (5)$$

A. Lower Bound

Because the capacity takes the maximum value over all input distributions, a lower bound of capacity can be derived using the independent and uniform distributed (i.u.d.) input.

Theorem 2. *The capacity of the DSEC can be lower bounded as*

$$C_{DSEC} \geq \left((1 - p_d) \log K - H(p_d) - \frac{1 - p_d}{1 - p_i} H(p_i) \right)^+. \quad (6)$$

Here $H(p) = -p \log p - (1 - p) \log(1 - p)$, which denotes the binary entropy function with the probability p .

Proof. To derive an achievable bound for the DSEC, we consider the situation that the input distribution is i.u.d.. The mutual information can be written as

$$\begin{aligned} I(X^n; Y^{N(n)}) &= I(X^n; Y^{N(n)}, D^n) - I(X^n; D^n | Y^{N(n)}) \\ &= H(X^n) - H(X^n | Y^{N(n)}, D^n) - H(D^n | Y^{N(n)}) + H(D^n | X^n, Y^{N(n)}). \end{aligned} \quad (7)$$

We now evaluate the four elements of the above equation.

- When X^n is i.u.d., we have $H(X^n) = nH(X) = n \log K$.
- $H(X^n | D^n, Y^{N(n)})$ can be written as

$$\begin{aligned} &H(X^n | D^n, Y^{N(n)}) \\ &= \sum_{i=1}^n H(X_i | X_{i-1}, X_{i-2}, \dots, X_1, D^n, Y^{N(n)}) \\ &= \sum_{i=1}^n \left(Pr[D_i = 0] H(X_i | D_i = 0, D^n, Y^{N(n)}) + Pr[D_i \neq 0] H(X_i | D_i \neq 0, D^n, Y^{N(n)}) \right) \\ &= n \cdot p_d \log K. \end{aligned} \quad (8)$$

The last equality holds because given D^n and $Y^{N(n)}$, if $D_i \neq 0$, the value of X_i will be fixed, and if $D_i = 0$, the value of X_i is unknown and X^n are i.u.d..

- Since the conditional action reduces entropy, we have

$$\begin{aligned} H(D^n | Y^{N(n)}) &\leq nH(D) \\ &= n \left(-p_d \log p_d - \sum_{r=0}^{\infty} p_t p_i^r \log p_t p_i^r \right) \end{aligned}$$

$$= n \left(H(p_d) + \frac{1-p_d}{1-p_i} H(p_i) \right). \quad (9)$$

- We have $H(D^n|X^n, Y^{N(n)}) \geq 0$ because of the non-negativity of entropy.

Therefore, when the input X_i are i.u.d., the mutual information

$$I(X^n|Y^{N(n)}) \geq n \left(\log K - H(p_d) - \frac{1-p_d}{1-p_i} H(p_i) - p_d \log K \right). \quad (10)$$

Then the capacity of DSEC can be lower bounded taking the i.u.d. input,

$$C_{DSEC} \geq \lim_{n \rightarrow \infty} \frac{I(X^n; Y^{N(n)})}{n} \geq \left((1-p_d) \log K - H(p_d) - \frac{1-p_d}{1-p_i} H(p_i) \right)^+. \quad (11)$$

□

Note that the above result implies the following for the binary channel,

$$C_{SEC} \geq \left((1-p_d) - H(p_d) - \frac{1-p_d}{1-p_i} H(p_i) \right)^+, \quad (12)$$

which is same as the result in [24].

B. Upper Bound

Indeed, to derive the capacity, the input should take over all distributions. We derive the general and theoretical upper bound of the DSEC due to the lack of methods for determining the optimal input distribution.

Theorem 3. *The capacity of the DSEC can be upper bounded as*

$$C_{DSEC} < (1-p_d) \log K + \frac{K-1}{K} (1-p_d)^2 p_i (1-p_i) \log(1-p_d)(1-p_i). \quad (13)$$

Note that if the value of the right side of the inequality is less than or equal to zero, the $C_{DSEC} = 0$.

Proof. The mutual information $I(X^n; Y^{N(n)})$, can be decomposed as

$$I(X^n; Y^{N(n)}) = I(X^n; Y^{N(n)}, D^n) - I(X^n; D^n | Y^{N(n)}). \quad (14)$$

Here $I(X^n; Y^{N(n)}, D^n)$ can be regarded as the mutual information of a channel such that the auxiliary sequence is also observed. With D^n , each $\vec{y}_i$ could be discerned from $Y^{N(n)}$. Plus, if no deletion happens,

the input symbol x_i could be uniquely determined from $\vec{y}_i$. Thus, $I(X^n; Y^{N(n)}, D^n)$ is equivalent to the mutual information of the memoryless erasure channel, where $p(Y^{N(n)}, D^n | X^n) = \prod_{i=1}^n p(\vec{Y}_i, D_i | X_i)$.

$$\begin{aligned}
I(X^n; Y^{N(n)}, D^n) &= H(Y^{N(n)}, D^n) - H(Y^{N(n)}, D^n | X^n) \\
&\leq \sum_{i=1}^n H(\vec{Y}_i, D_i) - \sum_{i=1}^n H(\vec{Y}_i, D_i | X_i) \\
&= \sum_{i=1}^n (H(\vec{Y}_i) - H(\vec{Y}_i | X_i)) \\
&= \sum_{i=1}^n I(X_i; \vec{Y}_i) \\
&\leq n(1 - p_d) \log K,
\end{aligned} \tag{15}$$

where $\vec{Y}_i$ denotes the output vector of X_i . When X_i is uniformly distributed for $i = 1, 2, \dots, n$, $H(X_i) = \log K$, the last equal sign is met.

Besides, we have

$$I(X^n; D^n | Y^{N(n)}) = H(D^n | Y^{N(n)}) - H(D^n | X^n, Y^{N(n)}), \tag{16}$$

where $H(D^n | Y^{N(n)}) \rightarrow H(D^n)$ as $n \rightarrow \infty$ because the only information obtained from $Y^{N(n)}$ about D^n is the sum of elements in D^n , which vanishes as $n \rightarrow \infty$. Consider $H(D^n | X^n, Y^{N(n)})$, because D varies from 0 to infinite, and the insertion can be an arbitrary sequence, the information obtained from $(X^n, Y^{N(n)})$ about D^n is restricted, and can be expanded as

$$\begin{aligned}
&H(D^n | X^n, Y^{N(n)}) \\
&= \sum_{i=1}^n H(D_i | X_i, Y_{N_D^n(i-1)+1}^n) \\
&= \sum_{i=1}^n \left(Pr[X_i = Y_{N_D^n(i-1)+1}^n] H(D_i | X_i = Y_{N_D^n(i-1)+1}^n) \right. \\
&\quad \left. + Pr[X_i \neq Y_{N_D^n(i-1)+1}^n] H(D_i | X_i \neq Y_{N_D^n(i-1)+1}^n) \right).
\end{aligned} \tag{17}$$

To simplify, we consider the first state,

$$H(D_1 | X_1 \neq Y_1) \leq H(D_1) + p_t \log p_t. \tag{18}$$

When $X_1 \neq Y_1$, we can see that the first symbol was not directly transmitted, where $p_t \log p_t < 0$ denotes

the reduction in the uncertainty of D_1 due to the knowledge of the $(X_1 \neq Y_1)$. And when the first symbol was inserted or deleted while the following symbol not same as it, there will have $X_1 \neq Y_1$. Let q_0 denote the probability such that $q_0 = \Pr[X_1 \neq X_2]$. We have,

$$\begin{aligned} \Pr[X_1 \neq Y_1] &\geq (p_i p_t + p_i^2 p_t + \dots) \frac{K-1}{K} + p_d q_0 \\ &= (1 - p_d) p_i \frac{K-1}{K} + p_d q_0. \end{aligned} \quad (19)$$

Hence, we have

$$H(D^n | X^n, Y^{N(n)}) \leq nH(D) + n \left(\frac{K-1}{K} (1 - p_d) p_i + p_d q_0 \right) p_t \log p_t. \quad (20)$$

However, due to the uncertainty of the optimal input distribution, we can only take $q_0 \neq 0$ as the result that the input can not be continuously identical symbols. Then the mutual information in (16) is upper bounded as

$$\lim_{n \rightarrow \infty} \frac{I(X^n; D^n | Y^{N(n)})}{n} > -\frac{K-1}{K} (1 - p_d) p_i p_t \log p_t. \quad (21)$$

Therefore, combine (15) and (21) the upper bound of the capacity can be limited as

$$C_{DSEC} < (1 - p_d) \log K + \frac{K-1}{K} (1 - p_d) p_i p_t \log p_t. \quad (22)$$

Here, if the value of the right side of the inequality is less than or equal to zero, the $C_{DSEC} = 0$. $\square$

Note that $\log(1 - p_d)(1 - p_i) < 0$. The above result yields a theoretical capacity upper bound for the DSEC that improves the trivial $(1 - p_d) \log K$ upper bound. This theorem shows that if there are any synchronization errors in transmission, it is impossible to achieve the capacity equal to the erasure channel.

C. Remarks

Recall the proof of the capacity bounds of the DSEC, the only inaccurate parameter when fixing the input distribution is $H(D^n | X^n, Y^{N(n)})$, which is the uncertainty of transmission tracks. Intuitively, when transmitting $X^n = (0, 0, \dots, 0)$, if the received is $Y^{n-1} = (0, 0, \dots, 0)$, the deletion may occur in every position. But when transmitting $X^n = (0, 1, 0, 1, \dots, 0, 1)$, if a deletion occurs, we can easily locate it between two consecutive identical elements. However, because the practical error types include deletion

and multiple insertions, the possible transmission tracks may be diverse, which makes it more difficult for us to find the real track. Therefore, for $H(D^n|X^n, Y^{N(n)})$ to be smaller, we expect that the codewords have special structures, which should be a trade-off with the uniformity of the input. This idea gives us some inspiration: add markers in the codewords may be more effective for decoding.

Besides, based on the Chain rule, the entropy $H(X^n) = \sum_{i=1}^n H(X_i|X_{i-1}, X_{i-2}, \dots, X_1)$. Recall that in theorem 2, we choose the i.u.d. input codewords $X^n = (X_1, X_2, \dots, X_n)$, so $H(X^n) = nH(X)$. Since the DSEC has memory, it is likely that a codeword structure with memory may perform better than an uniform distributed input. Indeed, the first-order Markov distributed input has tighter capacity lower bounds for the deletion channel [13],[25], and the sticky channel [26], which both are special cases of the DSEC. It gives us some inspiration: the convolutional codes may perform better.

IV. CAPACITY BOUNDS OF THE DNSEC

In this section, we characterize the capacity bounds of the DNSEC. The DNSEC has two classes of errors that are synchronization errors and additive errors. The synchronization errors have memory, but the additive errors do not. Hence, we separate these two type errors and consider the DNSEC as a cascaded channel of a DSEC and a DMC (Fig. 3).

Fig. 3. The DNSEC channel is cascaded by a DSEC and a DMC.

Take the input sequence of the DNSEC as $X^n = (X_1, X_2, \dots, X_n)$, and the output as $Z^{N(n)}$. Take the intermediate output of the DSEC as $Y^{N(n)}$, which is not observed in the DNSEC. Let $Y^{N(n)}|_I$ denote a subsequence of $Y^{N(n)}$ that is composed of the successfully transmitted symbols in X^n , where I signifies the indices of these symbols in $Y^{N(n)}$. And $\mathbb{E}(|Y^{N(n)}|_I) = (1 - p_d)n$. Suppose the additive error type of the DMC is $E^{|I|} = (E_1, E_2, \dots, E_{|I|}) \in \mathcal{K}^{|I|}$, and the error probability

$$Pr[E = e] = \begin{cases} 1 - p_s, & e = 0; \\ \frac{1}{K-1}p_s, & 1 \leq e \leq K-1. \end{cases} \quad (23)$$

Hence the output of the DMC can be written as $Z^{N(n)}|_I = Y^{N(n)}|_I \oplus E^{|I|}$, and $Z^{N(n)}|_{I^c} = Y^{N(n)}|_{I^c}$. In this model, the transmission of the DNSEC can be written as $X^n \rightarrow Y^{N(n)} \rightarrow Z^{N(n)}$. The transition probability satisfies,

$$p(\vec{z}^{N(n)}|\vec{x}^n) = \sum_{\vec{y}^{N(n)}} p(\vec{y}^{N(n)}|\vec{x}^n)p(\vec{z}^{N(n)}|\vec{y}^{N(n)}), \quad (24)$$

for any $\vec{x}^n \in X^n, \vec{y}^{N(n)} \in Y^{N(n)}, \vec{z}^{N(n)} \in Z^{N(n)}$.

A. Lower Bound

We now derive the lower bound of the DNSEC considering i.u.d. input codebooks as in theorem 2.

Theorem 4. *The capacity of DNSEC can be lower bounded as*

$$C_{DNSEC} \geq \left(1 - p_d\right) \left(\log K - p_s \log(K - 1) - H(p_s) \right) - H(p_d) - \frac{1 - p_d}{1 - p_i} H(p_i) \Big)^+. \quad (25)$$

Proof. The mutual information $I(X^n; Z^{N(n)})$ can be expanded to $H(X^n) - H(X^n|Z^{N(n)})$, we have

$$\begin{aligned} H(X^n|Z^{N(n)}) &= H(Y^{N(n)}|Z^{N(n)}) + H(X^n|Y^{N(n)}, Z^{N(n)}) - H(Y^{N(n)}|X^n, Z^{N(n)}) \\ &\leq H(Y^{N(n)}|Z^{N(n)}) + H(X^n|Y^{N(n)}), \end{aligned} \quad (26)$$

where the last inequality holds from the fact that $X^n \rightarrow Y^{N(n)} \rightarrow Z^{N(n)}$ is a Markov chain which suggests that X^n is independent of $Z^{N(n)}$ given $Y^{N(n)}$. Consider the situation that the input is i.u.d. as in section III, the output $Y^{N(n)}|_I$ of the DSEC is also i.u.d.. Knowing that the additive errors are symmetric, the conditional entropy $H(Y^{N(n)}|Z^{N(n)})$ can be written as

$$\begin{aligned} H(Y^{N(n)}|Z^{N(n)}) &= \sum_{i=1}^{N(n)} H(Y_i|Z_i) \\ &= \sum_{i \in I} H(Y_i|Z_i) + \sum_{i \in I^c} H(Y_i|Z_i) \\ &= n(1 - p_d)H(E) \\ &= n(1 - p_d) \left(- (1 - p_s) \log(1 - p_s) - (K - 1) \cdot \frac{1}{K - 1} p_s \log \frac{1}{K - 1} p_s \right) \\ &= n(1 - p_d) (H(p_s) + p_s \log(K - 1)). \end{aligned} \quad (27)$$

Therefore, the lower bound of the DNSEC can be derived. According to (26),

$$I(X^n; Z^{N(n)}) = H(X^n) - H(X^n|Z^{N(n)})$$

$$\begin{aligned}
&\geq H(X^n) - H(X^n|Y^{N(n)}) - H(Y^{N(n)}|Z^{N(n)}) \\
&= I(X^n; Y^{N(n)}) - H(Y^{N(n)}|Z^{N(n)}),
\end{aligned} \tag{28}$$

where $I(X^n; Y^{N(n)})$ is shown in theorem 2 when the input is i.u.d.. Hence, we have

$$C_{DNSEC} \geq \left((1 - p_d) (\log K - p_s \log(K - 1) - H(p_s)) - H(p_d) - \frac{1 - p_d}{1 - p_i} H(p_i) \right)^+. \tag{29}$$

□

B. Upper Bound

Theorem 5. *The capacity of DNSEC can be upper bounded as*

$$\begin{aligned}
C_{DNSEC} &< (1 - p_d) \left(\log K - H(p_s) - p_s \log(K - 1) \right) \\
&\quad + \frac{K - 1}{K} (1 - p_d)^2 p_i (1 - p_i) \log(1 - p_d) (1 - p_i).
\end{aligned} \tag{30}$$

If the value of the right side of the inequality is less than or equal to zero, $C_{DNSEC} = 0$.

Proof. Consider the DNSEC with the input $X^n \in \mathcal{K}^n$ and the output $Z^{N(n)}$, then its mutual information can be written as

$$I(X^n|Z^{N(n)}) = I(X^n; Z^{N(n)}, D^n) - I(X^n; D^n|Z^{N(n)}), \tag{31}$$

where D^n denotes the transmission track of X^n in DSEC. For the first part, $X^n \rightarrow (Z^{N(n)}, D^n)$ is a DMC with erasure errors and additive errors, we have

$$I(X^n; Z^{N(n)}, D^n) = I(X^n; Y^{N(n)}, D^n) - I(X^n; Y^{N(n)}, D^n|Z^{N(n)}, D^n), \tag{32}$$

where $I(X^n; Y^{N(n)}, D^n)$ denotes the mutual information of the DSEC, and $I(X^n; Y^{N(n)}, D^n|Z^{N(n)}, D^n)$ can be expanded as

$$I(X^n; Y^{N(n)}|Z^{N(n)}, D^n) = H(Y^{N(n)}|Z^{N(n)}, D^n) - H(Y^{N(n)}|X^n, Z^{N(n)}, D^n), \tag{33}$$

where

$$H(Y^{N(n)}|Z^{N(n)}, D^n) = n(1 - p_d) (H(p_s) + p_s \log(K - 1))$$

as in (27), $H(Y^{N(n)}|X^n, Z^{N(n)}, D^n) = 0$ is the result that $Y^{N(n)}$ can be determined by $(X^n, Z^{N(n)}, D^n)$. Hence, we have

$$I(X^n; Z^{N(n)}, D^n) < n(1 - p_d) \left(\log K - H(p_s) - p_s \log(K - 1) \right) + n \frac{K - 1}{K} (1 - p_d)^2 p_i (1 - p_i) \log(1 - p_d)(1 - p_i). \quad (34)$$

And for the second part, same like (16),

$$I(X^n; D^n | Z^{N(n)}) = H(D^n | Z^{N(n)}) - H(D^n | X^n, Z^{N(n)}) \geq 0. \quad (35)$$

Because of the variety of error types, the certainty of getting D^n from $(X^n, Z^{N(n)})$ is much limited, as a result, $H(D^n | X^n, Z^{N(n)})$ will approach $H(D^n)$ as $n \rightarrow \infty$. Hence, the capacity can be upper bounded as

$$C_{DNSEC} \leq \lim_{n \rightarrow \infty} \max_{P_{X^n}} \frac{I(X^n; Z^{N(n)})}{n} < (1 - p_d) \left(\log K - H(p_s) - p_s \log(K - 1) \right) + \frac{K - 1}{K} (1 - p_d)^2 p_i (1 - p_i) \log(1 - p_d)(1 - p_i). \quad (36)$$

Here, if the value of the right side of the inequality is less than or equal to zero, $C_{DNSEC} = 0$. $\square$

V. CAPACITY BOUNDS OF THE CONCATENATED DNA STORAGE CHANNEL

We proceed to evaluate the capacity bounds of the cascaded DNA storage channel. Let n and m denote the length and quantity of DNA strands respectively. Let q be the probability that a given strand is never sampled. When the sampled strands are error-free, the capacity of the binary SSC is provided in [7],

$$C_{SSC} = (1 - q)(1 - 1/\beta), \quad (37)$$

where $\beta = \lim_{m \rightarrow \infty} \frac{n}{\log m}$ denotes the loss rate of index due to the disordered fashion of the sequencing. They also generalized this conclusion to the scenario of a Noisy Shuffling Channel (*NSC*) which is cascaded by a Symmetric Discrete Memoryless Channel (*SDMC*) and a SSC [7], and have

$$C_{NSC} = (1 - q)(C_{SDMC} - 1/\beta). \quad (38)$$

Similar to [7], we assume that a strand is either sampled once with probability $1 - q$ or not sampled with probability q , which can be identified as a simple Bernoulli distribution. We cascade such the SSC

with the DNSEC to estimate the theoretical limits of the DNA storage channel (Fig.4).

Fig. 4. The DNA storage channel is cascaded by a DNSEC and a SSC.

Theorem 6. For the DNA storage channel, let n and m denote the length and quantity of DNA strands respectively, and $\beta = \lim_{m,n \rightarrow \infty} \frac{n}{\log m}$, the storage capacity can be bounded as

$$\begin{aligned} & \left((1-q) \left((1-p_d)(2-H(p_s) - p_s \log 3) - H(p_d) - \frac{1-p_d}{1-p_i} H(p_i) - \frac{1}{2\beta} \right) \right)^+ \\ & \leq C_{DNA} < \\ & (1-q) \left((1-p_d)(2-H(p_s) - p_s \log 3) + \frac{3}{4}(1-p_d)^2 p_i (1-p_i) \log(1-p_d)(1-p_i) - \frac{1}{\beta} \right), \end{aligned} \quad (39)$$

where p_d , p_i and p_s denote the deletion, insertion, and substitution probability of individual DNA molecules respectively, and q denotes the probability that a given strand is never sampled. Note that if the value of the right side of the inequality is less than or equal to zero, $C_{DNA} = 0$.

Proof. For the DNA storage channel cascaded by an SSC and a DNSEC, a typical coding systems employs an interleaved structure. The outer code is designed to recover the message within dropout strands, and the inner code is used to correct the errors in individual DNA molecules.

To find the lower bound, we consider such a concatenated coding scheme. We apply an erasure correcting code with m codewords and rate $(1-q)$ as the outer code, where each codeword has n quaternary symbols. In [7], it is proved that a simple index-based coding scheme is optimal for the outer channel. Hence, $\beta' = \lim_{m,n \rightarrow \infty} \frac{n}{\log_4 m} = 2\beta$ denotes the redundancy rate to store DNA strand IDs. The number of symbols stored valid information in each codeword is restricted to $(nR_{i.u.d.} - \log_4 m)$, where $R_{i.u.d.}$ is the achievable rate of the inner code as described in theorem 4. Hence, a lower bound of the capacity of the DNA storage

channel can be determined as

$$\lim_{m,n \rightarrow \infty} (1-q)m \frac{nR_{i.u.d.} - \log_4 m}{mn} = (1-q)(R_{i.u.d.} - 1/2\beta). \quad (40)$$

To find the upper bound, consider a set of codewords for the DNA storage channel with rate R and vanishing decoding error probability. Let $X^{mn} = (X_1^n, X_2^n, \dots, X_m^n)$ denote the input which was chosen from the uniform distribution on $\{1, 2, \dots, 2^{mnR}\}$, and let $Z^{MN} = (Z_1^{N(n)}, Z_2^{N(n)}, \dots, Z_M^{N(n)})$ denote the corresponding output with unknown lengths and disordered indices. Note that $N(n)$ is a dummy function that varies for different sequences. And we have $\lim_{m \rightarrow \infty} \mathbb{E}(M)/m = (1-q)$. Here, let S_i denote the correct correspondence between Z^{MN} and X^{mn} , such that $Z_i^{N(n)}$ is transmitted from $X_{S_i}^n$. However, we only can obtain an estimated index function $\hat{S}^M$ based on decoded strand IDs, but it is not perfectly accurate due to synchronization errors and additive errors in transmission. According to Fano's inequality, we have $H(X^{mn}|Z^{MN}) \leq 1 + P_e^{(m)}mnR$, where $P_e^{(m)}$ is the decoding error probability. Hence, we have

$$\begin{aligned} mnR &= H(X^{mn}) = I(X^{mn}; Z^{MN}) + H(X^{mn}|Z^{MN}) \\ &\leq H(Z^{MN}) - H(Z^{MN}, S^M|X^{mn}) + H(S^M|Z^{MN}, X^{mn}) + 1 + P_e^{(m)}mnR \\ &= H(Z^{MN}) + H(Z^{MN}|X^{mn}, S^M) - H(S^M|X^{mn}) + H(S^M|Z^{MN}, X^{mn}) + 1 + P_e^{(m)}mnR \\ &= \sum_{i=1}^M \left(H(Z_i^{N(n)}) - H(Z_i^{N(n)}|X_{S_i}^n) \right) - H(S^M) + H(S^M|Z^{MN}, X^{mn}) + 1 + P_e^{(m)}mnR \\ &= \sum_{i=1}^M I(Z_i^{N(n)}; X^n) - H(S^M) + H(S^M|Z^{MN}, X^{mn}) + 1 + P_e^{(m)}mnR. \end{aligned} \quad (41)$$

Here, S^M is a sequence with distinct elements which are chosen uniformly from $\{1, 2, \dots, m\}^M$. Hence, we have

$$H(S^M) = \log \frac{m!}{(m-M)!} = M \log m + (m-M) \log \frac{m}{m-M} + o(m), \quad (42)$$

according to the Stirling approximation. And by Jensen's inequality,

$$\sum_{i=1}^m \Pr(M=i)(m-i) \log \frac{m}{m-i} \leq (1-q)m \log 1/q = o(m). \quad (43)$$

Next, $H(S^M|Z^{MN}, X^{mn})$ can be regarded as the uncertainty of the corresponding relationship between Z^{MN} and X^{mn} . Let $g : Z^{MN} \rightarrow X^{mn}$ denote the decoding function, and $g(Z_i^N) = X_{\hat{S}_i}^n$. We have, $\Pr[S_i \neq$

$\hat{S}_i] = P_e^{(m)}$. According to Fano's inequality,

$$\begin{aligned}
& H(S^M | Z^{MN}, X^{mn}) \\
& \leq \sum_{i=1}^M \left(H(S_i | \hat{S}_i) + H(P_e^{(m)}) \right) \\
& = \sum_{i=1}^M \left(\Pr[\hat{S}_i = S_i] H(S_i | \hat{S}_i) + \Pr[\hat{S}_i \neq S_i] H(S_i | \hat{S}_i) + H(P_e^{(m)}) \right) \\
& \leq P_e^{(m)} M \log(m - (1 - P_e^{(m)})M) + MH(P_e^{(m)}), \tag{44}
\end{aligned}$$

where and the last inequality holds from the fact that the maximum number of choices of S_i is limited to the number of X^n that are never correctly matched, which is no more than $(m - (1 - P_e^{(m)})M)$.

Therefore, we have

$$\begin{aligned}
mnR(1 - P_e^{(m)}) & \leq \sum_{i=1}^M I(X^n; Z^{N(n)}) - M \log m + MH(P_e^{(m)}) \\
& \quad + P_e^{(m)} M \log(m - (1 - P_e^{(m)})M) + o(m) + 1. \tag{45}
\end{aligned}$$

Since we consider the situation that $P_e^{(m)} \rightarrow 0$ and $m, n \rightarrow \infty$, so we have $R \leq (1 - q)(R_{inner} - 1/\beta)$, where $R_{inner} = \lim_{n \rightarrow \infty} \frac{1}{n} I(X^n; Z^{N(n)})$ can be upper bounded as in theorem 5.

□

VI. DISCUSSION

In this paper, we for the first time constructed a channel model termed DNSEC to study the information transmission over single DNA strands with synchronization errors as well as additive errors. We derived the lower and the upper capacity bounds of the DNSEC based on information theory. The DNSEC could be considered as a more generalized synchronization channel. By taking $p_i = p_s = 0$, it is a deletion channel. By taking $p_d = p_s = 0$, it is a sticky channel. When limiting $K = 2$, the binary synchronization error channel can be obtained, the lower capacity bound in theorem 2 is equal to the results in [24]. Then we provided a unified framework to describe the DNA storage channel by cascading the SSC and the DNSEC. By analyzing the theoretical limits of this cascaded channel, we bounded the storage capacity of DNA-based storage systems.

As described in section II, error rates in DNA synthesis and sequencing are much higher than in silicon-based storage systems. In particular, a significant fraction of them are synchronization errors which are

less investigated in previous studies, so how to model and analyze them is still an underexamined area. The questions that have been widely discussed around the DNA-based storage technology are whether it is a high-efficient and sufficient storage medium, and where is the limit of storage density of DNA. As shown in Fig.5(a), when the average error rate per position is 0.6% as observed in the current DNA storage systems, the inner storage capacity of DNA molecules is impressive, which suggests that DNA is a promising storage medium with high storage capacity. Novel-designed coding structures are desired to achieve the bounds of the DNSEC capacity.

(a) Different error rates ($p_d = p_i = p_s = p_r$)

(b) Different error rates ($p_d = p_i = p_s = p_r$)

Fig. 5. Performance of the capacity bounds versus error rates. (a) shows the magnified portion of (b) with the DNA mutation error probability $p_d = p_i = p_s = p_r \leq 0.01$.

Several interesting questions remain open about the DNSEC. The central question is the single-letter characterization of its capacity. However, it seems quite challenging as in other synchronization channels. A more practical direction is to develop tighter upper and lower capacity bounds. As shown in Fig.5(b), the gap between lower and upper bounds are still large, which means that i.i.d. input distribution is far from optimal. In particular, the general SECs which allow deletion and multiple insertions per symbol are more complicate than deletion channels and sticky channels. Besides, design techniques for synchronization error-correction codes are desired for the accurate information extraction in DNA-based storage systems.

ACKNOWLEDGMENT

We would like to thank our colleagues at the Center for Applied Mathematics, Tianjin University: Dr. Alan J.X. Guo, Dr. Huaming Wu, and Guanjin Qu who helped us a lot. This work is supported by the National Key R&D Program of China (grant No. 2020YFA0712102) and the National Natural Science Foundation of China (grant no. 12001401).

REFERENCES

- [1] G. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in dna," *Science (New York, N.Y.)*, vol. 337, p. 1628, 08 2012.
- [2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. Leproust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized dna," *Nature*, vol. 494, 01 2013.
- [3] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on dna in silica with error-correcting codes," *Angew Chem Int Ed Engl*, vol. 54, no. 8, pp. 2552–2555, 2015.
- [4] L. Organick, S. D. Ang, Y. J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, and B. Nguyen, "Random access in large-scale dna data storage," *Nature Biotechnology*, vol. 36, no. 3, 2018.
- [5] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in dna," *Science*, vol. 337, no. 6102, p. 1628, 2012.
- [6] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the dna data storage channel," *Scientific Reports*, vol. 9, no. 1, p. 9663, 2019. [Online]. Available: <https://doi.org/10.1038/s41598-019-45832-6>
- [7] I. Shomorony and R. Heckel, "Dna-based storage: Models and fundamental limits," *IEEE Transactions on Information Theory*, vol. PP, pp. 1–1, 02 2021.
- [8] H. Mercier, V. K. Bhargava, and V. Tarokh, "A survey of error-correcting codes for channels with symbol synchronization errors," *IEEE Communications Surveys Tutorials*, vol. 12, no. 1, pp. 87–96, 2010.
- [9] M. Cheraghchi and J. Ribeiro, "An overview of capacity results for synchronization channels," *IEEE Transactions on Information Theory*, vol. PP, pp. 1–1, 05 2020.
- [10] S. Shin, R. Heckel, and I. Shomorony, "Capacity of the erasure shuffling channel," pp. 8841–8845, 05 2020.
- [11] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "An upper bound on the capacity of the dna storage channel," in *2019 IEEE Information Theory Workshop (ITW)*, 2019, pp. 1–5.

- [12] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakohi, "Achieving the capacity of the dna storage channel," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
- [13] S. Diggavi and M. Grossglauser, "On information transmission over a finite buffer channel," in *IEEE*, 2006, pp. 1226–1237.
- [14] R. GALLAGER, "Sequential decoding for binary channels with noise and synchronization errors," p. 1, 10 1961.
- [15] E. Drinea and M. Mitzenmacher, "On lower bounds for the capacity of deletion channels," *IEEE Transactions on Information Theory*, vol. 52, pp. 4648–4657, 2006.
- [16] —, "Improved lower bounds for the capacity of i.i.d. deletion and duplication channels," *IEEE Transactions on Information Theory*, vol. 53, no. 8, pp. 2693–2714, 2007.
- [17] J. D. Ullman, "On the capabilities of codes to correct synchronization errors," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 95–105, 1967.
- [18] S. Diggavi, M. Mitzenmacher, and H. D. Pfister, "Capacity upper bounds for the deletion channel," in *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*, 2007.
- [19] M. Sack, N. Kretschy, B. Rohm, V. Somoza, and M. M. Somoza, "Simultaneous light-directed synthesis of mirror-image microarrays in a photochemical reaction cell with flare suppression," *Analytical Chemistry*, vol. 85, no. 18, 2013.
- [20] M. Karl, C. John, C. Marcelo, C. James, S. Dominic, G. Andrey, J. A. Leonetti, W. Wei, F. M. Rossi, and A. G. St?Ver, "Electrochemically generated acid and its containment to 100 micron reaction areas for the production of dna microarrays," *Plos One*, vol. 1, no. 1, p. e34, 2006.
- [21] S. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free dna-based data storage," *Scientific Reports*, vol. 7, no. 1, p. 5011, 2017.
- [22] Shannon and E. C., "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, no. 4, pp. 623–656, 1948.
- [23] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," *Probl.peredachi Inf*, vol. 3, no. 4, p. 18ndash;36, 1967.
- [24] A. R. Iyengar, P. H. Siegel, and J. K. Wolf, "Modeling and information rates for synchronization error channels," pp. 380–384, 2011.
- [25] C. Torre, M. Miguel, J. Vázquez, M. Ramírez, S. Barrena, F. Hernández, J. Mariño, L. Lassaletta, and J. B. Ortega, "Improved lower bounds for the capacity of i.i.d. deletion and duplication channels," *IEEE Press*, 2007.
- [26] M. Mitzenmacher, "Capacity bounds for sticky channels," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 72–77, 2008.