The Secure Storage Capacity of a DNA Wiretap Channel Model

Praneeth Kumar Vippathalla and Navin Kashyap

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834344

***Abstract*—In this paper, we propose a strategy for making DNA-based data storage information-theoretically secure through the use of wiretap channel coding. This motivates us to extend the shuffling-sampling channel model of Shomorony and Heckel (2021) to include a wiretapper. Our main result is a characterization of the *secure storage capacity* of our DNA wiretap channel model, which is the maximum rate at which data can be stored within a pool of DNA molecules so as to be reliably retrieved by an authorized party (Bob), while ensuring that an unauthorized party (Eve) gets almost no information from her observations. Furthermore, our proof of achievability shows that index-based wiretap channel coding schemes are optimal.**

I. Introduction

In DNA-based storage, raw data (e.g., text) is converted into sequences over an alphabet consisting of the building blocks of DNA, namely, the four nucleotide bases, adenine (A), guanine (G), thymine (T), and cytosine (C). The sequences over the DNA alphabet are then physically realized by artificially synthesizing DNA molecules (called oligonucleotides, or oligos, in short) corresponding to the string of A, G, T, C letters forming the sequences. These synthetic DNA molecules (oligos) can be mass-produced and replicated to make many thousands of copies and the resulting pool of oligos can be stored away in a controlled environment. At the time of data retrieval, sequencing technology is used to determine the A-G-T-C sequence forming each oligo from the pool. Prior to sequencing, the pool of oligos is subjected to several cycles of Polymerase Chain Reaction (PCR) amplification. In each cycle of PCR, each molecule in the pool is replicated ("amplified") by a factor of 1.6–1.8. The PCR amplification process requires knowledge of short initial segments (prefixes) and final segments (suffixes) of the oligos to be amplified. This knowledge is used to design the required primers to initiate PCR amplification. After PCR, a small amount of material from the amplified pool is passed through a sequencing platform [1] that randomly samples and sequences (by patching together "reads" of a relatively short length) the molecules from the pool. The raw data is then retrieved from these sequences.

While DNA-based data storage technology provides a reliable solution for long-term data storage, there often can arise security issues. Suppose Alice wants to store using DNA-based data storage some sensitive information which is to be retrieved later by a trusted party, Bob. A solution to this problem is that Alice uses a private key K, which is shared with Bob, to one-time pad her information W and then store it into a pool of synthetic DNA oligonucleotides. But a significant drawback of this scheme is that the size of the key K should be comparable to the size of information W, which is not practically feasible. Another simple solution, proposed by Clelland *et al*. [2], is to design the oligos that encode W in such a way that the keystring K can be converted to the specific primers needed for initiating PCR-based amplification of these oligos. Then, these oligos can be synthesized in small amounts (low copy numbers), and then hidden within an ocean of "background" or "junk" DNA. The background DNA could, for example, be fragments from the human genome. A vial containing the composite pool of information-bearing and background DNA is stored in a DNA-storage repository.

Since Bob knows K, given the vial containing the composite oligo pool, he can provide the primers needed to selectively amplify only the information-bearing oligos using PCR. After several cycles of PCR, the copy numbers of these oligos in the post-PCR sample become large enough that when the sample is fed into a sequencing platform, each of these oligos get a large number of reads. After filtering out the reads that come from the background DNA (this is easily possible if the background DNA is made up of human genome fragments), the remaining reads correspond to the information-bearing oligos. These reads can be assembled using standard sequence assembly algorithms, after which their information content, W, can be recovered by Bob.

Eve, on the other hand, does not know K, so is unable to selectively amplify the information-bearing oligos. Instead, her only option is to amplify the entire composite oligo pool, including background DNA, using whole-library PCR. (In whole-library PCR, known adapters are non-selectively ligated onto the oligos in the pool, and the complementary sequences of these adapters are used as primers to initiate PCR.) However, this process does not discriminate between information-bearing oligos and background DNA, so it does not significantly change the proportion of information-bearing molecules to background DNA. Since background DNA still constitutes the overwhelming majority of molecules in the amplified pool, it is highly unlikely that the information-bearing oligos will get sufficiently many reads to get reliably sequenced. This protects the data W from being reliably recovered by Eve.

One issue is that Eve's protocol will allow her to partially recover the data W, which may be undesirable. This is

Praneeth Kumar V. (praneethv@iisc.ac.in) and N. Kashyap (nkashyap@iisc.ac.in) are with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012. This work was supported in part by a grant, IE/RERE-19-0514, from the Indian Institute of Science.

Authorized licensed use limited to: R.L. Kashyap Memorial Library Indian Institute of Science Bengaluru. Downloaded on September 07,2022 at 10:40:09 UTC from IEEE Xplore. Restrictions apply.

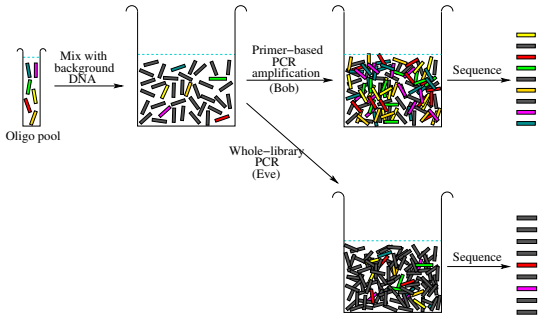

Fig. 1. The basic strategy of secure DNA-based data storage.

possible even after just one round of sequencing, but if her budget allows for multiple rounds of sequencing, then she can use the sequences recovered from the first round to determine some of the primers. She can then use these primers for selective amplification in the next round of sequencing to boost her chances of getting more information about W. The more the rounds of sequencing she is allowed, the better she can do.

A. Our proposed strategy: Using coding to boost security

What we propose is the use of coding to greatly (and cheaply) improve upon the basic scheme above. As observed above, the use of the shared key K as primers gives Bob an advantage over Eve. In information-theoretic parlance, we say that the use of primers creates a "channel" from Alice to Bob that is much less noisy (much more reliable) than the "channel" from Alice to Eve. In the area of information-theoretic security, this situation is called a *wiretap channel* [3], and the design of coding schemes for such a channel is a well-investigated research topic – see e.g., [4], [5], [6]. These coding schemes greatly boost any advantage that Bob has over Eve, however small it may initially be. To clarify, the aim of a wiretap channel coding scheme is to enable Alice to encode the data W into a codeword X (in our case, this will be realized as a pool of oligos) such that the difference in noise statistics between the Alice-Bob and Alice-Eve channels can be exploited, as envisioned below:

(a) When X passes through the channel from Alice to Bob, what Bob receives is a slightly noisy version of X, which we denote by Y, from which he is able to recover the data W highly reliably;

(b) but when X passes through the channel from Alice to Eve, Eve observes a significantly noisier version Z, from which she gets almost *no information* about the data W.

The proportion, ρ, of information-bearing molecules in the composite pool is a knob that we can use to control the noise statistics of Eve's channel — see [7, Appendix A] for a brief discussion of how ρ affects Eve's probability of oligo erasure. It is desirable to keep ρ small (10^{-3} or lower) so as to put Eve at a significant disadvantage. But it cannot be so small that the primers used to initiate PCR fail to find a sufficient quantity of information-bearing molecules to amplify, causing even Bob's protocol to fail.

The advantages of our approach over the basic approach are two-fold:

(i) Our scheme should be able to tolerate higher ratios ρ than the basic scheme of Clelland et al. [2], as it is enough to choose a ρ that gives a minor statistical advantage for Bob over Eve. Higher ratios ρ will allow Bob's protocol to work much more reliably.

(ii) Wiretap channel coding is supposed to ensure that Eve gets essentially no information about the data W, *no matter what strategy she uses* to recover W from her observations Z.

B. Main contribution

We build on the recent pioneering work of Shomorony and Heckel [8] that models and determines the fundamental limits of DNA-based data storage (but without any considerations of security). In this work, data W is encoded in the form of an oligo pool, which is viewed as a multiset X, each element of which is a sequence of length L over the DNA alphabet $\Sigma = \{A, C, G, T\}$. From the multiset X, a multiset of N sequences is randomly drawn according to some fixed probability distribution P. The P distribution encapsulates the randomness inherent in the processes of oligo synthesis, PCR amplification and sequencing. The resulting multiset, Y, of sequences is observed by the decoder, which attempts to retrieve W from Y. This channel is referred to as a *(noise-free) shuffling-sampling channel* with distribution P. Shomorony and Heckel [8] studied the *storage capacity* of this model, defined to be the maximum rate at which data W can be stored in an oligo pool X, and reliably retrieved from the observed multiset Y. Their study in fact extends to a *noisy* model wherein the N sequences sampled from X may further be corrupted by insertion, deletion and substitution errors affecting the individual letters making up each sequence.

We extend the noise-free Shomorony-Heckel model above by introducing an additional noise-free shuffling-sampling channel with distribution Q for the unauthorized party (Eve). Bob's ability to selectively amplify information-bearing oligos using his knowledge of primers allows us to assume that the probability, q_0, that a particular oligo is not seen by Eve is larger than the corresponding probability, p_0, for Bob. We study the *secure storage capacity*, C_s, of the noise-free shuffling-sampling model, which we define to be the maximum rate at which data W can be stored in an oligo pool X, and reliably retrieved from Bob's observation Y, while ensuring that Eve gets (almost) no information about W from her observations. We completely characterize C_s, giving an expression for it that depends only on q_0 and p_0.

II. DNA Storage Wiretap Channel Model

Let M denote the number of DNA molecules in the oligo pool, and L denote the length of each molecule. Though $\{A, C, G, T\}$ is the alphabet of DNA coding, we work with $\Sigma := \{0, 1\}$ for the sake of simplicity. However, in the case of general alphabet Σ, the results will involve an extra $\log_2 |\Sigma|$ term. Fix a constant $\beta := \lim_{M\to\infty} \frac{L}{\log M} > 1$; throughout this

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bangalore. Downloaded on September 07,2022 at 10:40:09 UTC from IEEE Xplore. Restrictions apply.

paper, all logarithms are to the base 2. Let W be a uniform random variable taking values in $\mathcal{W} \triangleq \{1, 2, \ldots, 2^{MLR}\}$ for some $R \geq 0$. Alice encodes (maps) the message W into M DNA molecules, each of length L, which is denoted by a multiset $X^{ML} = \{X_1^L, \ldots, X_M^L\}$. The stored X^{ML} is amplified and sequenced to recover the message W. A fundamental model that captures the cumulative effect of these processes without errors is the *noise-free shuffling sampling channel* with some distribution $(\pi_0, \pi_1, \ldots)$ [8]. This channel randomly permutes (shuffles) the order of the L-length molecules, and independently outputs each molecule $n \geq 0$ times with probability π_n. The output of this channel is also a multiset. We can use this model for the amplification and synthesis at both Bob and Eve's side. Hence Bob and Eve observe multisets $Y^{N_{\mathrm{m}}L} = \{Y_1^L, \ldots, Y_{N_{\mathrm{m}}}^L\}$ and $Z^{N_{\mathrm{w}}L} = \{Z_1^L, \ldots, Z_{N_{\mathrm{w}}}^L\}$, respectively, which are obtained by passing the input X^{ML} through two independent noise-free shuffling sampling channels with distributions $P = (p_0, p_1, \ldots)$ and $Q = (q_0, q_1, \ldots)$, respectively. See Fig. 2. The quantities[1] N_{m} and N_{w} are random variables taking non-negative integer values, and Y_i^L, Z_j^L are elements of Σ^L. The goal of Alice is to store a message that can only be recovered by Bob while keeping Eve in ignorance of it.

Formally, let $\phi : \mathcal{W} \longrightarrow \mathbb{N}^{\Sigma^L}$ be an encoding function (possibly a stochastic function) of Alice, and $\psi : \mathbb{N}^{\Sigma^L} \longrightarrow \mathcal{W} \cup \{\mathbf{e}\}$ be a decoding function of Bob, where $\mathbf{e}$ denotes an error, and $\mathbb{N}^{\Sigma^L}$ denotes the set of all multisets with finite cardinality over Σ^L. We say that a secure message rate R is achievable if there exists a sequence of pairs of encoding and decoding functions $\{(\phi, \psi)\}_{M=1}^{\infty}$ that satisfy Bob's recoverability condition,

$$\mathbb{P}\{\psi(Y^{N_{\mathrm{m}}L}) \neq W\} \to 0, \tag{1}$$

and the (strong) secrecy condition,

$$I(W; Z^{N_{\mathrm{w}}L}) \to 0 \tag{2}$$

as $M \to \infty$. The *secure storage capacity*, C_s, is defined by

$$C_s \triangleq \sup\{R : R \text{ is achievable}\} \tag{3}$$

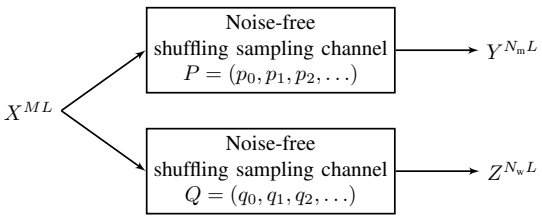

Fig. 2. DNA storage wiretap channel model

In the $q_0 = 1$ case, where none of the molecules are sampled by Eve's channel, the secure storage capacity is nothing but the storage capacity of Bob's channel [8].

A useful fact about multisets is that they can be uniquely identified with frequency vectors. Given a multiset A whose elements are from a finite set $\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\}$, we denote by $\mathbf{f}_A = [\mathrm{f}_A(a_1), \ldots, \mathrm{f}_A(a_{|\mathcal{A}|})]$ the frequency vector corresponding to A. For $a_i \in \mathcal{A}$, the component $\mathrm{f}_A(a_i)$ counts the number of occurrences of a_i in the set A. In a slight abuse of notation, we use $\mathbf{f}_X$, $\mathbf{f}_Y$ and $\mathbf{f}_Z$ to denote the frequency (random) vectors corresponding to the multisets $X^{ML}, Y^{N_{\mathrm{m}}L}$ and $Z^{N_{\mathrm{w}}L}$, respectively, over Σ^L. As frequency vectors and multisets are interchangeable, the Markov chain $W - X^{ML} - (Y^{N_{\mathrm{m}}L}, Z^{N_{\mathrm{w}}L})$ is equivalent to

$$W - \mathbf{f}_X - (\mathbf{f}_Y, \mathbf{f}_Z) \tag{4}$$

for any joint distribution that is induced by an encoding function. The components of a frequency vector will be typeset in regular font, for example f_X. Let $h(x)$ denote the binary entropy function, i.e, $h(x) = -x \log x - (1-x)\log(1-x)$, for $x \in (0, 1)$. The n-dimensional probability simplex, denoted by Δ^n, is defined as $\Delta^n \triangleq \left\{(x_0, \ldots, x_n) \in \mathbb{R}^{n+1} : \sum_{i=0}^{n} x_i = 1\right\}$.

III. Secure Storage Capacity

Our main result is the following expression for the secure storage capacity of our model.

Theorem 1. *For a DNA storage wiretap channel with $q_0 \geq p_0$,*

$$C_s = \left(1 - \frac{1}{\beta}\right)(q_0 - p_0). \tag{5}$$

Since the $q_0 = 1$ case corresponds to the storage capacity of Bob's channel, we get the result [8, Theorem 1] as a corollary of the above theorem.

Corollary 1 ([8], Theorem 1)**.** *The storage capacity of a noise-free shuffling-sampling channel with the distribution $P = (p_0, p_1, \ldots)$ is equal to $\left(1 - \frac{1}{\beta}\right)(1 - p_0)$.*

The remainder of this section is devoted to a proof of Theorem 1, which we divide into an achievability part and a converse part.

A. Achievability part

Alice encodes a message by using distinct indices for each of the M sequences. The initial segment of length $\log M$ of a sequence contains the index, and the rest of the sequence is used to encode the message. However, since the index is only used to order the molecules, the rate is scaled by a factor of $\left(1 - \frac{1}{\beta}\right)$. The technique of indexing converts a noise-free shuffling-sampling channel with distribution $(\pi_0, \pi_1, \ldots)$ into a block-erasure channel (acting on a block of length L) with erasure probability π_0. As a result, both Bob and Eve's channels are equivalent to erasure channels with erasure probabilities $\epsilon_{\mathrm{m}} = p_0$ and $\epsilon_{\mathrm{w}} = q_0$, respectively. So, for this wiretap channel, a rate of $\epsilon_{\mathrm{w}} - \epsilon_{\mathrm{m}} = q_0 - p_0$ is achievable by encoding a message securely using the random coding arguments of [3], [9]. Though these techniques give a scheme satisfying the weak secrecy condition, i.e., $\frac{1}{ML} I(W; Z^{N_{\mathrm{w}}L}) \to 0$, the resulting secrecy can be strengthened to satisfy (2) by using the ideas of [10]. Hence,

$$\left(1 - \frac{1}{\beta}\right)(q_0 - p_0) \leq C_s. \tag{6}$$

[1]The subscripts 'm' and 'w' stand for "main" and "wiretap", respectively.

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bengaluru. Downloaded on September 07,2022 at 10:40:09 UTC from IEEE Xplore. Restrictions apply.

B. Converse part

We prove the converse by considering a modified scenario. In this new scenario, Bob has access to a (genie-aided) side information S in addition to $Y^{N_{\mathrm{m}}L}$. As the side information can only increase the ability of Bob to recover the message, the capacity of this scenario is at least C_s. For Eve, we weaken her observations by providing her with a processed version of $Z^{N_{\mathrm{w}}L}$. Here, "processed version of $Z^{N_{\mathrm{w}}L}$" means that it is a random variable obtained by applying a stochastic function (independent of everything else) to $Z^{N_{\mathrm{w}}L}$. It is clear that weakening Eve can only make the secrecy capacity larger, so that

$$C_s \leq \hat{C}_s, \tag{7}$$

where $\hat{C}_s$ is the secure storage capacity of the model with a genie-aided Bob and a weaker Eve.

The side information S we give to Bob is the same as that considered in the proof of [8, Theorem 1]: If Y_i^L and Y_j^L, $i \neq j$ are two identical L-length molecules, then S distinguishes whether they were sampled from the same input molecule X_k^L or from two identical input molecules $X_{k_1}^L$ and $X_{k_2}^L$, $k_1 \neq k_2$. Using $(S, Y^{N_{\mathrm{m}}L})$, Bob can compute the multiset $\hat{Y}^{M_{\mathrm{m}}L} \subseteq \{X_1^L, \ldots, X_M^L\}$ that contains all the molecules of X^{ML} that were sampled at least once. Here the random variable M_{m} denotes the cardinality of the multiset. Let $\mathbf{f}_{\hat{Y}}$ denote the frequency vector corresponding to $\hat{Y}^{M_{\mathrm{m}}L}$ over Σ^L. Note that the distribution of $(S, Y^{N_{\mathrm{m}}L})$ depends on $\mathbf{f}_X$ only through $\mathbf{f}_{\hat{Y}}$, which implies the Markov chain

$$W - \mathbf{f}_X - \mathbf{f}_{\hat{Y}} - (S, Y^{N_{\mathrm{m}}L}). \tag{8}$$

Instead of $Z^{N_{\mathrm{w}}L}$, Eve has only access to $\hat{Z}^{M_{\mathrm{w}}L}$, which is the set of all distinct L-length molecules in Σ^L that appear in $Z^{N_{\mathrm{w}}L}$. Let $\mathbf{f}_{\hat{Z}}$ denote the frequency vector corresponding to $\hat{Z}^{M_{\mathrm{w}}L}$ over Σ^L. The entries of the frequency vector are $\mathrm{f}_{\hat{Z}}(a_i) = \mathbb{1}\{\mathrm{f}_Z(a_i) > 0\}$ for $a_i \in \Sigma^L$, which indicates whether an L-length molecule appears in the multiset $Z^{N_{\mathrm{w}}L}$ or not. By (4), we have $W - \mathbf{f}_X - \mathbf{f}_Z - \mathbf{f}_{\hat{Z}}$. While there are other choices for Eve that provide a reasonable estimate for X^{ML} through $Z^{N_{\mathrm{w}}L}$ (maximum likelihood (ML) estimate of $\mathbf{f}_X$ based on $\mathbf{f}_Z$ is one such choice), we choose $\hat{Z}^{M_{\mathrm{w}}L}$ for the purpose of simpler analysis.

Let us derive an upper bound on the secure storage capacity $\hat{C}_s$ of the new scenario, where Bob has $(S, Y^{N_{\mathrm{m}}L})$ and Eve has $\hat{Z}^{M_{\mathrm{w}}L}$. Suppose that R is an achievable secure message rate for the new scenario, i.e., there exists a sequence of pair of encoding and decoding functions $\{(\phi, \psi)\}_{M=1}^{\infty}$ that satisfy Bob's recoverability condition, $\mathbb{P}\{\psi(Y^{N_{\mathrm{m}}L}, S) \neq W\} \to 0$, and the secrecy (strong) condition, $I(W; \hat{Z}^{M_{\mathrm{w}}L}) \to 0$ as $M \to \infty$. Then R can be upper bounded as follows.

$$\begin{aligned}
MLR &= H(W) \\
&\overset{(a)}{\leq} I(W; Y^{N_{\mathrm{m}}L}, S) + MLR\delta_M + 1 \\
&\overset{(b)}{\leq} I(W; Y^{N_{\mathrm{m}}L}, S) - I(W; \hat{Z}^{N_{\mathrm{w}}L}) + \delta'_M + MLR\delta_M + 1 \\
&\overset{(c)}{\leq} I(W; \mathbf{f}_{\hat{Y}}) - I(W; \mathbf{f}_{\hat{Z}}) + \delta'_M + MLR\delta_M + 1
\end{aligned}$$

where (a) is because of the inequality $H(W \mid Y^{N_{\mathrm{m}}L}, S) \leq 1 + MLR\delta_M$ for $\delta_M \to 0$, which is a consequence of Bob's recoverability condition (1) and Fano's inequality, (b) follows from Eve's secrecy condition (2) where $\delta'_M \to 0$, and (c) holds because of the data processing inequality $I(W; Y^{N_{\mathrm{m}}L}, S) \leq I(W; \mathbf{f}_{\hat{Y}})$ for the Markov chain (8), and $I(W; \hat{Z}^{N_{\mathrm{w}}L}) = I(W; \mathbf{f}_{\hat{Z}})$. We can rewrite the above upper bound on R as

$$(1 - \delta_M)R - \delta''_M \leq \frac{1}{ML}[I(W; \mathbf{f}_{\hat{Y}}) - I(W; \mathbf{f}_{\hat{Z}})] \tag{9}$$

where $\delta''_M = \frac{\delta'_M + 1}{ML} \to 0$, $\delta_M \to 0$ and $W - \mathbf{f}_X - (\mathbf{f}_{\hat{Y}}, \mathbf{f}_{\hat{Z}})$.

1) Degradation of the frequency vector channels: The term $I(W; \mathbf{f}_{\hat{Y}}) - I(W; \mathbf{f}_{\hat{Z}})$ in (9) depends on $P_{\mathbf{f}_{\hat{Y}}, \mathbf{f}_{\hat{Z}} | \mathbf{f}_X}$ only through the marginal distributions $P_{\mathbf{f}_{\hat{Y}} | \mathbf{f}_X}$ and $P_{\mathbf{f}_{\hat{Z}} | \mathbf{f}_X}$. Hence, with no loss in generality, we can work with a new coupled distribution $\tilde{P}_{\mathbf{f}_{\hat{Y}}, \mathbf{f}_{\hat{Z}} | \mathbf{f}_X}$ that has the same marginals as that of $P_{\mathbf{f}_{\hat{Y}}, \mathbf{f}_{\hat{Z}} | \mathbf{f}_X}$. For that, first note that the conditional joint distribution $P_{\mathbf{f}_{\hat{Y}}, \mathbf{f}_{\hat{Z}} | \mathbf{f}_X}$ is $P_{\mathbf{f}_{\hat{Y}} | \mathbf{f}_X} P_{\mathbf{f}_{\hat{Z}} | \mathbf{f}_X}$ because given $\mathbf{f}_X$, Bob's side information S depends only on the Bob's noise-free shuffling-sampling channel which is independent of Eve's channel. Furthermore, we can decompose the channel between $\mathbf{f}_X$ and $\mathbf{f}_{\hat{Y}}$(or $\mathbf{f}_{\hat{Z}}$) into $|\Sigma^L|$ i.i.d. channels, one for each of the components of the frequency vectors, because the sampling process is independent across all the DNA molecules (see Fig. 3). Since X^{ML} contains M molecules, the input frequency vector is constrained by $\sum_{i=1}^{|\Sigma^L|} \mathrm{f}_X(a_i) = M$. Let $P_{\mathrm{f}_{\hat{Y}} \mathrm{f}_{\hat{Z}} | \mathrm{f}_X} = P_{\mathrm{f}_{\hat{Y}} | \mathrm{f}_X} P_{\mathrm{f}_{\hat{Z}} | \mathrm{f}_X}$ denote the channel transition probability between the components $\mathrm{f}_X(a_i)$ and $(\mathrm{f}_{\hat{Y}}(a_i), \mathrm{f}_{\hat{Z}}(a_i))$, for $a_i \in \Sigma^L$. Since $\mathrm{f}_{\hat{Y}}$ is a sum of f_X number of independent Ber(p_0) random variables and $\mathrm{f}_{\hat{Z}} = \mathbb{1}\{\mathrm{f}_Z > 0\}$, the channel transition probabilities are given by

$$P_{\mathrm{f}_{\hat{Y}} | \mathrm{f}_X}(j \mid i) = \begin{cases} \binom{i}{j} p_0^{i-j}(1 - p_0)^j, & \text{if } j \leq i \text{ and} \\ 0, & \text{otherwise} \end{cases}$$

and

$$P_{\mathrm{f}_{\hat{Z}} | \mathrm{f}_X}(j \mid i) = \begin{cases} q_0^i, & \text{if } j = 0 \text{ and} \\ 1 - q_0^i, & \text{if } j = 1 \end{cases}$$

where the alphabets for f_X, $\mathrm{f}_{\hat{Y}}$ and $\mathrm{f}_{\hat{Z}}$ are $\{0, 1, \ldots, M\}$, $\{0, 1, \ldots, M\}$ and $\{0, 1\}$, respectively.

Lemma 1. *If $q_0 \geq p_0$, then the channel $P_{\mathrm{f}_{\hat{Z}} | \mathrm{f}_X}$ is a degraded version of $P_{\mathrm{f}_{\hat{Y}} | \mathrm{f}_X}$, i.e., there exists a channel Q such that*

$$P_{\mathrm{f}_{\hat{Z}} | \mathrm{f}_X}(j|i) = \sum_{k=0}^{M} Q(j|k) P_{\mathrm{f}_{\hat{Y}} | \mathrm{f}_X}(k|i)$$

for all $i \in \{0, 1, \ldots, M\}$ and $j \in \{0, 1\}$.

A proof of the above lemma can be found in the extended version of this paper [7]. By using Lemma 1, we consider the coupled distribution $\tilde{P}_{\mathrm{f}_{\hat{Y}} \mathrm{f}_{\hat{Z}} | \mathrm{f}_X} = P_{\mathrm{f}_{\hat{Y}} | \mathrm{f}_X} Q$

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bangalore. Downloaded on September 07,2022 at 10:40:09 UTC from IEEE Xplore. Restrictions apply.

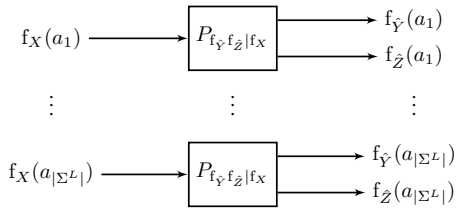

Fig. 3. The channel between $\mathbf{f}_X$ and $(\mathbf{f}_{\hat{Y}}, \mathbf{f}_{\hat{Z}})$ is decomposed into $|\Sigma^L|$ i.i.d. channels between the components with transition probability $P_{\mathrm{f}_{\hat{Y}}\mathrm{f}_{\hat{Z}}|\mathrm{f}_X}$. The input $\mathbf{f}_X$ is constrained by $\sum_{i=1}^{|\Sigma^L|} \mathrm{f}_X(a_i) = M$.

for the component channels with the marginals $P_{\mathrm{f}_{\hat{Z}}|\mathrm{f}_X}$ and $P_{\mathrm{f}_{\hat{Y}}|\mathrm{f}_X}$. This type of coupling for all the component channels yield a joint distribution that satisfy the Markov chain $W - \mathbf{f}_X - \mathbf{f}_{\hat{Y}} - \mathbf{f}_{\hat{Z}}$. So, we can write $I(W;\mathbf{f}_{\hat{Y}}) - I(W;\mathbf{f}_{\hat{Z}}) = I(W;\mathbf{f}_{\hat{Y}} \mid \mathbf{f}_{\hat{Z}}) \le I(\mathbf{f}_X;\mathbf{f}_{\hat{Y}} \mid \mathbf{f}_{\hat{Z}})$.

2) An upper bound in terms of the component channel mutual information: For $\sum_{i=1}^{|\Sigma^L|} \mathrm{f}_X(a_i) = M$, carrying on from (9), we have

$$
\begin{aligned}
&(1-\delta_M)R - \delta''_M \\
&\quad\le \frac{1}{ML}[I(W;\mathbf{f}_{\hat{Y}}) - I(W;\mathbf{f}_{\hat{Z}})] \\
&\quad\le \frac{1}{ML} I(\mathbf{f}_X;\mathbf{f}_{\hat{Y}} \mid \mathbf{f}_{\hat{Z}}) \\
&\quad= \frac{1}{ML}[H(\mathbf{f}_{\hat{Y}} \mid \mathbf{f}_{\hat{Z}}) - H(\mathbf{f}_{\hat{Y}} \mid \mathbf{f}_X, \mathbf{f}_{\hat{Z}})] \\
&\quad= \frac{1}{ML} \sum_{i=1}^{|\Sigma^L|} \big[H(\mathrm{f}_{\hat{Y}}(a_i) \mid \mathbf{f}_{\hat{Z}}, \mathrm{f}_{\hat{Y}}(a_1), \dots \mathrm{f}_{\hat{Y}}(a_{i-1})) \\
&\qquad\quad -H(\mathrm{f}_{\hat{Y}}(a_i) \mid \mathbf{f}_X, \mathbf{f}_{\hat{Z}}, \mathrm{f}_{\hat{Y}}(a_1), \dots \mathrm{f}_{\hat{Y}}(a_{i-1}))\big] \\
&\quad\overset{(a)}{\le} \frac{1}{ML} \sum_{i=1}^{|\Sigma^L|} \big[H(\mathrm{f}_{\hat{Y}}(a_i) \mid \mathrm{f}_{\hat{Z}}(a_i)) \\
&\qquad\quad -H(\mathrm{f}_{\hat{Y}}(a_i) \mid \mathrm{f}_X(a_i), \mathrm{f}_{\hat{Z}}(a_i))\big] \qquad (10) \\
&\quad= \frac{1}{ML} \sum_{i=1}^{|\Sigma^L|} I(\mathrm{f}_X(a_i);\mathrm{f}_{\hat{Y}}(a_i) \mid \mathrm{f}_{\hat{Z}}(a_i)) \\
&\quad= \frac{|\Sigma^L|}{ML} \sum_{i=1}^{|\Sigma^L|} \frac{1}{|\Sigma^L|} I(\mathrm{f}_X(a_i);\mathrm{f}_{\hat{Y}}(a_i) \mid \mathrm{f}_{\hat{Z}}(a_i)) \qquad (11) \\
&\quad\overset{(b)}{\le} \frac{|\Sigma^L|}{ML} I(\mathrm{f}_X;\mathrm{f}_{\hat{Y}} \mid \mathrm{f}_{\hat{Z}}) \qquad (12) \\
&\quad\le \frac{|\Sigma^L|}{ML} \sup f(m_0, \dots, m_M). \qquad (13)
\end{aligned}
$$

where (a) follows from the fact that conditioning reduces entropy and that $\mathrm{f}_{\hat{Y}}(a_i)$ is conditionally independent of $\mathbf{f}_X, \mathbf{f}_{\hat{Z}}, \mathrm{f}_{\hat{Y}}(a_1), \dots \mathrm{f}_{\hat{Y}}(a_{i-1})$ given $\mathrm{f}_X(a_i)$ and $\mathrm{f}_{\hat{Z}}(a_i)$, and (b) is a consequence of the fact that for a Markov chain $X - Y - Z$, $I(X;Y|Z)$ is a concave function in the distribution of X [11, Lemma 1]. The summation in (11) is a convex combination of conditional mutual information terms $I(\mathrm{f}_X(a_i);\mathrm{f}_{\hat{Y}}(a_i) \mid \mathrm{f}_{\hat{Z}}(a_i))$ evaluated with respect to an input distribution $(m_0(i), \dots, m_M(i)) \in \Delta^M$ where $m_j(i)$ denotes the probability (induced by the encoder) that the component $\mathrm{f}_X(a_i)$ equals j. One can verify that $\sum_{i=1}^{|\Sigma^L|} \sum_{j=0}^{M} j m_j(i) = M$ under the input constraint $\sum_{i=1}^{|\Sigma^L|} \mathrm{f}_X(a_i) = M$. In (12), $I(\mathrm{f}_X;\mathrm{f}_{\hat{Y}} \mid \mathrm{f}_{\hat{Z}})$ is evaluated at the input distribution $(m_0, \dots, m_M) = \frac{1}{|\Sigma^L|}\Big(\sum_{i=1}^{|\Sigma^L|} m_0(i), \dots, \sum_{i=1}^{|\Sigma^L|} m_M(i)\Big)$, which satisfies the constraint $\sum_{j=1}^{M} j m_j = \frac{M}{|\Sigma^L|}$, and the supremum of $f(m_0, \dots, m_M) := I(\mathrm{f}_X;\mathrm{f}_{\hat{Y}} \mid \mathrm{f}_{\hat{Z}})$ in (13) is over such input distributions. By taking limits on both sides, we get the bound

$$\hat{C}_s \le \liminf_{M\to\infty} \frac{|\Sigma^L|}{ML} \sup f(m_0, \dots, m_M) \qquad (14)$$

where, again, the supremum is over $(m_0, \dots, m_M) \in \Delta^M$ subject to $\sum_{j=1}^{M} j m_j = \frac{M}{|\Sigma^L|}$. For sufficiently large M, $\sup f(m_0, \dots, m_M)$ is bounded above by

$$h\left(\frac{M}{|\Sigma^L|}(1-p_0)\right) - h\left(\frac{M}{|\Sigma^L|}(1-q_0)\right) + o\left(\frac{ML}{|\Sigma^L|}\right).$$

A proof of this bound is given in the extended version of this paper [7]. By using this bound in (14), we obtain

$$
\begin{aligned}
\hat{C}_s &\le \frac{\beta-1}{\beta} \liminf_{x\to 0} \frac{h((1-p_0)x) - h((1-q_0)x)}{-x\log x} \\
&\overset{(c)}{=} \left(1 - \frac{1}{\beta}\right)(q_0 - p_0),
\end{aligned}
$$

where we set $x := \frac{M}{|\Sigma^L|}$ and use the fact that $\frac{-\log x}{L} = \frac{L - \log M}{L} \to 1 - \frac{1}{\beta}$ as $M \to \infty$, and the evaluation of the limit in (c) requires the use of L'Hôpital's rule (twice). Combining this with (7), we have $C_s \le \left(1 - \frac{1}{\beta}\right)(q_0 - p_0)$, which completes the proof of the converse part of Theorem 1.

IV. Discussion

The DNA storage wiretap channel model considered in this paper is motivated by the fact that differential knowledge of primers creates a statistical advantage for the authorized party, Bob, over the unauthorized party, Eve. The take-away message from our work is that by exploiting this advantage using wiretap channel coding schemes, we can obtain information-theoretically secure DNA-based storage within our model.

The crucial part of our characterization of the secure storage capacity of our model is the converse proof, which is based on analytically solving an optimization problem. An alternative proof, which is along the lines of that given for the converse part of Theorem 1 in [8], is provided in [7, Appendix C] for the special case when Eve's sampling distribution is Bernoulli $Q = (q_0, q_1)$. We were unable to find a way of extending this argument directly to more general distributions Q.

In future work, we intend to consider the DNA wiretap channel model with noisy shuffling-sampling channels as its components.

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bengaluru. Downloaded on September 07,2022 at 10:40:09 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, Jun. 2016.

[2] C. T. Clelland, V. Risca, and C. Bancroft, "Hiding messages in DNA microdots," *Nature*, vol. 399, no. 6736, pp. 533–534, Jun. 1999.

[3] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975.

[4] I. Csiszar and J. Korner, "Broadcast channels with confidential messages," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 339–348, May 1978.

[5] M. Bloch and J. Barros, *Physical-Layer Security: From Information Theory to Security Engineering*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[6] A. Thangaraj, S. Dihidar, A. R. Calderbank, S. W. McLaughlin, and J.-M. Merolla, "Applications of LDPC codes to the wiretap channel," *IEEE Trans. Inf. Theory*, vol. 53, no. 8, pp. 2933–2945, Aug. 2007.

[7] P. K. Vippathalla and N. Kashyap, "The secure storage capacity of a DNA wiretap channel model," 2022. [Online]. Available: https://arxiv.org/abs/2201.05995

[8] 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.

[9] L. H. Ozarow and A. D. Wyner, "Wire-tap channel II," *AT&T Bell Lab. Tech. J.*, vol. 63, no. 10, pp. 2135–2157, Dec. 1984.

[10] U. Maurer and S. Wolf, "Information-theoretic key agreement: From weak to strong secrecy for free," in *Proc. EUROCRYPT (Lecture Notes in Computer Science)*, vol. 1807. Springer-Verlag, 2000, pp. 351–368.

[11] S. Leung-Yan-Cheong, "On a special class of wiretap channels (corresp.)," *IEEE Trans. Inf. Theory*, vol. 23, no. 5, pp. 625–627, Sep. 1977.

Authorized licensed use limited to: J.R.D. Tata Memorial Library Indian Institute of Science Bengaluru. Downloaded on September 07,2022 at 10:40:09 UTC from IEEE Xplore. Restrictions apply.