Achieving the Capacity of a DNA Storage Channel with Linear Coding Schemes

Kel Levick
University of Illinois, Urbana-Champaign
Urbana, IL, USA
klevick2@illinois.edu

Reinhard Heckel
Technical University of Munich
Munich, Germany
reinhard.heckel@tum.de

Ilan Shomorony
University of Illinois, Urbana-Champaign
Urbana, IL, USA
ilans@illinois.edu

2022 56th Annual Conference on Information Sciences and Systems (CISS) | 978-1-6654-1796-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CISS53076.2022.9751151

***Abstract*—Due to the redundant nature of DNA synthesis and sequencing technologies, a basic model for a DNA storage system is a multi-draw "shuffling-sampling" channel. In this model, a random number of noisy copies of each sequence is observed at the channel output. Recent works have characterized the capacity of such a DNA storage channel under different noise and sequencing models, relying on sophisticated typicality-based approaches for the achievability. Here, we consider a multi-draw DNA storage channel in the setting of noise corruption by a binary erasure channel. We show that, in this setting, the capacity is achieved by linear coding schemes. This leads to a considerably simpler derivation of the capacity expression of a multi-draw DNA storage channel than existing results in the literature.**

***Index Terms*—DNA storage, channel capacity, linear codes**

I. Introduction

Due to its longevity and high information density, DNA has drawn growing interest in its potential for archival data storage. Thanks to recent advancements in DNA sequencing (reading) and synthesizing (writing), this idea is becoming practically viable, and several groups have recently demonstrated working DNA storage systems [1–7]. In these systems, data is usually stored on short DNA molecules (a few hundred nucleotides). The synthesis process is usually redundant and produces a large number of copies of each molecule. At the time of reading, state-of-the-art sequencing technologies access this information, which corresponds to randomly sampling and reading sequences from the (redundant) DNA pool. Additionally, sequencing and synthesis may introduce errors to each sequence, most commonly in the form of insertions, substitutions, and deletions.

A natural mathematical model for DNA storage that accounts for these constraints is as follows: Data is stored onto M sequences, each of length L. This can be thought of as a single codeword of length ML, broken into M pieces of equal length. During sequencing, N sequences are randomly drawn from this set. Since the synthesis process is redundant and produces many copies of each molecule, and since the sequencing process is often preceded by Polymerase Chain Reaction (PCR), which effectively amplifies the number of copies of each molecule in the pool, several of the sequenced molecules correspond to the same original input sequence. However, because the synthesis and sequencing processes are noisy, the observed sequences are corrupted by *distinct noise patterns*. Therefore, an end-to-end model for DNA storage that captures this output sequence redundancy is the noisy shuffling-sampling channel with *multi-draws* [8], shown in Figure 1.

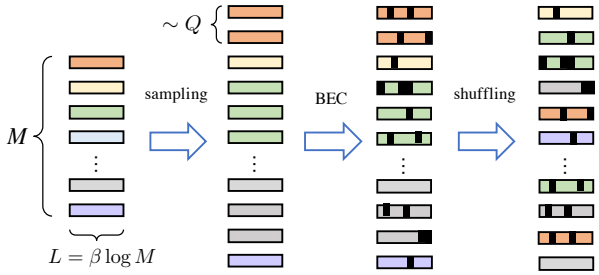

Fig. 1. The BEC multi-draw DNA Storage channel.

First, each of the M input strings are amplified a random number of times. The resulting molecules are independently corrupted by a noisy channel and shuffled out of order. Notice that some molecules may be sampled zero times, corresponding to the case where they are not sequenced at all. The decoder must then, without any knowledge of which molecules were sampled, reconstruct the original stored message using the sampled sequences at the channel output.

In this multi-draw setting, a clustering problem arises from the need to identify which of the output sequences correspond to the same input sequence. If we are able to correctly cluster the output, we can combine the sequences in each cluster to correct errors and decode the stored message using the (partially) error-corrected sequences.

The capacity of the multi-draw DNA storage channel with binary symmetric noise has been established using typicality based arguments [8, 9] and for general discrete memoryless channels based on the method of types [10]. While these arguments are interesting and novel, the analysis of the achievable rate is sophisticated and does not provide direct intuition for the resulting capacity expression. Moreover, the resulting decoding algorithms are computationally intractable, raising the question of whether simpler approaches such as linear codes can achieve the channel capacity.

Motivated by the fact that the capacity of a binary erasure channel (BEC) can be straightforwardly achieved using linear codes, in this work we consider the multi-draw DNA storage channel with binary erasure noise. More precisely, we focus on

a multi-draw shuffling-sampling channel where each (binary) sequence is corrupted by a binary erasure channel with erasure probability p, and each input string is drawn a random Q number of times, where Q has the probability mass function $\Pr(Q = n) = q_n$, for $n = 0, 1, 2, \ldots$. As in previous works, we consider the asymptotic regime where $M \to \infty$ and the read length scales as $L = \beta \log M$. The simplicity of the BEC setting allows us to show that a (random) linear coding scheme achieves the capacity of this channel for a large set of parameters (p, β), illustrated in Figure 2.

Based on the proposed linear coding scheme, we show that, for the blue regime in Figure 2, the capacity is given by

$$C = (1 - q_0)(E[C_Q|Q \geq 1] - 1/\beta), \tag{1}$$

where C_n is the capacity of a multi-draw shuffling-sampling channel with exactly n draws of each input sequence. As it turns out, the capacity expression in (1) can be verified to be equivalent to the expression obtained in related works [8, 10]. However, the expression in (1) is more intuitive and directly recovers all previous DNA storage channel capacity results, including the original results for DNA storage channels with "one or none draws" ($q_0 + q_1 = 1$) [11], in which case $E[C_Q|Q \geq 1] = C_1$. Hence, we conjecture that (1) is the general capacity formula for an arbitrary multi-draw DNA storage channel.

A. Related literature

The information-theoretic analysis of DNA storage channels started in [12] with a noise-free shuffling-sampling channel, and was later extended to noisy shuffling-sampling channels [11, 13] by modeling the noise as a BSC, and considering a single-draw setting where strings are drawn either once with probability $1 - q_0$ or not at all with probability q_0.

The single-draw DNA storage channel with BEC noise was considered in [14]. The concept of *consistency*, where two strings x_1^L, x_2^L with erasures are said to be consistent if they agree on every non-erased position, is used to establish the capacity for a set of parameters (p, β). We will also make use of the notion of consistency to create consistency graphs from the output strings in Section III.

The multi-draw shuffling-sampling channel was first studied in [8, 9]. The capacity was characterized for a regime of (p, β) and for the case where the output strings are independently observed through a BSC. The achievability argument was based on a random codebook construction. The decoder performs a greedy-like clustering of the output strings, and then uses typicality decoding based on a new notion of typicality between a set of d output strings and an input string.

Capacity results for multi-draw DNA storage channels were recently generalized to arbitrary discrete memoryless channels [10]. A general achievability was provided based on the method of types and a general upper bound was developed by refining the approach used in previous works [8, 11].

The problem of clustering output strings was studied from a coding-theoretic standpoint in [15]. It was shown that "code-aware" clustering, i.e., a clustering algorithm that exploits knowledge of the codebook (as opposed to a code-oblivious clustering), can reduce the number of redundancy bits needed for correct decoding.

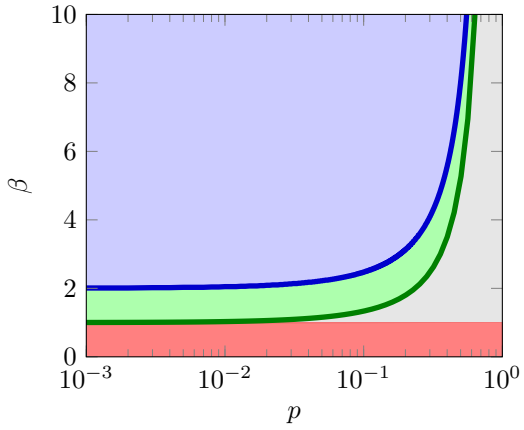

Fig. 2. The outer bound from (4) holds in the blue region, which corresponds to $\beta > 2/(1 - 2p + p^2)$. The inner bound from Theorem 2 holds above the green line, which corresponds to $\beta > -1/\log(1 - \frac{1}{2}(1 - p)^2)$. This characterizes the capacity of the BEC shuffling channel in the blue region as $(1 - q_0)(1 - p_{\text{eff}} - 1/\beta)$. The capacity in the red region (i.e., for $\beta < 1$) is 0 and it is unknown in the gray region.

II. Main Result

We consider the multi-draw DNA storage channel illustrated in Figure 1. The channel input is a length-ML binary string $X^{ML} = \left[X_1^L, X_2^L, \ldots, X_M^L\right]$, or, equivalently, M strings of length L concatenated to form a single string of length ML. Each of the M input strings is independently sampled Q times, where $\Pr(Q = n) = q_n$, for $n = 0, 1, \ldots$, yielding a set of N sequences. Each of these sequences is independently passed through an binary erasure channel with erasure probability p, and the final N sequences are shuffled out of order. We refer to the resulting end-to-end channel as the BEC multi-draw DNA storage channel.

As in previous works, we let $L = \beta \log M$, and consider the asymptotic regime $M \to \infty$. A rate R is said to be achievable if there exists a sequence of codes, each with 2^{MLR} codewords, and whose error probability goes to zero as $M \to \infty$. The channel capacity C is the supremum over all achievable rates.

We say that a code is a *linear coding scheme* if the 2^{MLR} length-ML codewords are all in the range of a $ML \times B$ binary generator matrix $\mathbf{G}$. Notice that we differentiate this from a *linear code*, in which case the set of codewords must be the entire range of $\mathbf{G}$. Our main result is the following.

Theorem 1. *For $\beta > 2/(1 - 2p + p^2)$, the capacity of the BEC multi-draw DNA storage channel is*

$$C = (1 - q_0)\left(E[C_{BEC,Q}|Q \geq 1] - 1/\beta\right), \tag{2}$$

and it can be achieved with a linear coding scheme. Here, $C_{BEC,n} = 1 - p^n$ is the capacity of a BEC with n draws.

A natural approach to deal with the output of a BEC multi-draw DNA storage channel is to first cluster output sequences based on consistency, and then combine all of the strings in each cluster into a "consensus" sequence, where the ith position of the length-L consensus sequence is the ith symbol from any of the strings of the cluster that is not erased.

If the clustering can be done successfully, this reduces the probability of erasure to p^n for an output cluster with n strings. Therefore, we expect that after consensus there will be $(1-q_0)M$ output strings, and for $n = 1, 2, \ldots$, there will be $q_n M$ consensus strings with erasure probability p^n. The resulting effective channel after the consensus step is similar to the single-draw BEC case as discussed in [14], but rather than a bit erasure probability of p across all strings, here the average effective erasure probability is given as

$$p_{\text{eff}} \triangleq E[p^Q | Q \geq 1] = \frac{\sum_{n=1}^{\infty} q_n p^n}{1 - q_0}. \tag{3}$$

Note that $E[C_{\text{BEC},Q} | Q \geq 1] = 1 - E[p^Q | Q \geq 1] = 1 - p_{\text{eff}}$. Since the capacity of BEC single-draw DNA storage channel [14] can be found to be

$$(1 - q_0)(1 - p - 1/\beta),$$

it is natural to conjecture that the capacity of the BEC multi-draw DNA storage channel is given by (2).

The converse to Theorem 1 can be found by considering a genie-aided argument where the genie reveals the true clusters, which can be used to find consensus sequences, and then using the result for the single-draw setting [14]. More formally, the converse can be obtained from the general DMC DNA storage channel capacity given in [10, Corollary 11]. Evaluating this result for the BEC yields

$$C = (1 - q_0)(1 - p_{\text{eff}} - 1/\beta) \tag{4}$$

for the regime $\beta > 2/(1 - 2p + p^2)$. This is represented by the blue area in Figure 2. Next, we show that for a larger set of parameters (p, β) (given by the green region), the capacity expression in (2) can be achieved with linear coding schemes.

III. Achievability via Linear Schemes

Motivated by the fact that linear codes achieve the capacity of the BEC [16], we show that linear coding schemes achieve the capacity of the BEC multi-draw DNA storage channel.

To construct a code of rate R, we first populate a random binary generator matrix $\mathbf{G}$ of size $ML \times B$, for some B to be determined, with i.i.d. Bernoulli(1/2) entries. We then generate 2^{MLR} length-B random binary vectors $\mathbf{t}_i$, also with i.i.d. Bernoulli(1/2) entries. The ith codeword of our random code, for $i = 1, \ldots, 2^{MLR}$, is constructed first by computing the product $\mathbf{G}\mathbf{t}_i$ over $\mathbb{F}_2$, and then by breaking the resulting codeword into M binary strings of length L. The channel input of M strings thus represents a single codeword.

We will use the following lemma throughout the proofs in this section.

Lemma 1. *Let $\mathbf{G}$ be an $ML \times B$ matrix with i.i.d.* Bernoulli$(1/2)$ *entries. Fix any $\delta \in (0,1)$ and a submatrix $\mathbf{G}'$ formed by an arbitrary set of $(1-\delta)B$ rows of $\mathbf{G}$. Then $\mathbf{G}'$ is full rank (over the finite field $\mathbb{F}_2$) with probability tending to 1 as $B \to \infty$.*

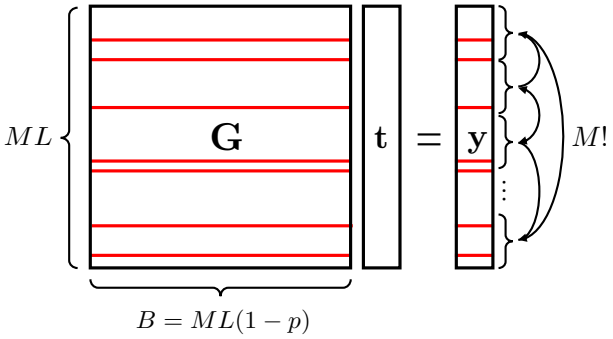

Fig. 3. In the setting where each sequence is observed exactly once at the output, there are $M!$ potential systems of equations, each with $ML(1-p)$ non-erased equations. Choosing our rate to be $\approx 1 - p - 1/\beta$ guarantees that, with high probability, only one of these systems will have a solution that corresponds to a codeword.

A. Single-draw case

We first illustrate the linear coding scheme by considering the case where each string is sampled exactly once, i.e., $q_1 = 1$. At the output of this system, $N = M$ strings are observed, which we wish to use to recover the stored message $\mathbf{t}_i$.

If the correct ordering of the M output strings were known, we would be able to concatenate them into a length-ML vector $\mathbf{y}$ and try to solve the system $\mathbf{G}\mathbf{t} = \mathbf{y}$ for $\mathbf{t}$. With erasure probability p, in expectation there are $ML(1-p)$ non-erased positions in $\mathbf{y}$, so a unique solution to this system exists with high probability as long as the number of remaining equations, which is roughly $ML(1-p)$, is greater than or equal to the number of variables B. Therefore, B can be set to $ML(1-p-\epsilon)$ for any $\epsilon > 0$ and all binary strings in $\{0,1\}^B$ may be used as message vectors.

However, the correct ordering of the output strings is not actually known, so we consider all $M!$ possible orderings. With each ordering, we have a distinct concatenated vector $\mathbf{y}$ with a p fraction of erasures, as well as a p fraction of useless equations in $\mathbf{G}$, as shown in Figure 3. From Lemma 1, if we set $B = ML(1-p-\epsilon)$, then the true ordering of remaining equations will have a unique solution.

In addition, we must have only one feasible ordering of the output strings; that is, out of the $M!$ possible systems of equations $\mathbf{G}\mathbf{t} = \mathbf{y}$, only one of them (the true one) should have a solution $\mathbf{t}$ that is one of the original message vectors $\mathbf{t}_i$, for $i = 1, \ldots, 2^{MLR}$. Assume without loss of generality that message 1 is sent (i.e., the codeword sent is $\mathbf{G}\mathbf{t}_1$). Since the messages are generated i.i.d. from $\{0,1\}^B$, by the union bound, the probability that there is a collision between the solution $\mathbf{t}$ of one of the $M! - 1$ incorrect systems and one of the other messages $\mathbf{t}_i$, $i = 2, \ldots, 2^{MLR}$ is at most

$$\begin{aligned}(M! - 1)2^{MLR}2^{-B} &< 2^{M \log M + MLR - B} \\ &= 2^{ML[R - (1-p-\epsilon-1/\beta)]}.\end{aligned}$$

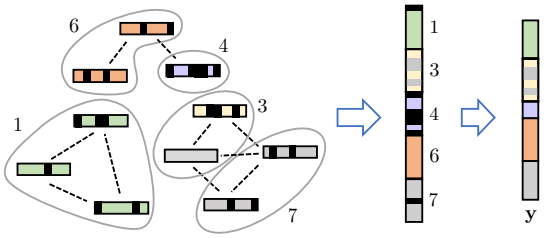

Fig. 4. The decoder first builds a consistency graph between all received sequences, and considers all possible partitions of the graph into cliques, where the total number of cliques is between $p_L M$ and $p_U M$. For each such valid clustering, the decoder considers all possible assignments of the indices $\{1, ..., M\}$ to the clusters and uses those indices to order the consensus sequences of each cluster to form the vector $\mathbf{y}$. After removing erased entries and the rows of $\mathbf{G}$ corresponding to erased/missing rows in $\mathbf{y}$, the system $\mathbf{Gt} = \mathbf{y}$ can be solved.

Therefore, by choosing ϵ arbitrarily small, we conclude that any rate $R < 1 - p - 1/\beta$ can be achieved with vanishingly small error probability.

B. Multi-draw case

We now consider the channel in Figure 1 with a general sampling distribution Q, i.e., each input string x_i^L is sampled $N_i \sim Q$ times. We now expect to see $E[N_1]M$ output strings, and we would like to cluster them into roughly $(1 - q_0)M$ clusters. This clustering task becomes easier under the BEC setting, as we can take advantage of the consistency of the strings. Notice that any two output strings that originated from the same input string are consistent. Therefore, given N output strings $x_1^L, \ldots, x_N^L$, one can construct an undirected graph with the strings as vertices and edges between any two consistent strings. We refer to this graph as a *consistency graph*. A clustering of these output strings is valid if it corresponds to a partition of $\{x_1^L, \ldots, x_N^L\}$ such that each group corresponds to a *clique* in the consistency graph.

With the probability of an input string not having an output cluster equal to q_0, in expectation there are $(1 - q_0)M$ output clusters. From Hoeffding's inequality, the probability that there are more than $p_U M := (1 - q_0 + \epsilon)M$ or fewer than $p_L M := (1 - q_0 - \epsilon)M$ output clusters can be bounded as

$$\Pr\left(|\#\text{ true output clusters} - (1 - q_0)M| > \epsilon M\right) < 2e^{-2M\epsilon^2},$$

which tends to 0 as $M \to \infty$ for any $\epsilon > 0$. The task of the decoder is then to cluster the N output strings into between $p_L M$ and $p_U M$ clusters for some small fixed ϵ; here, the decoder will consider all valid clusterings that create between $p_L M$ and $p_U M$ clusters. For each clustering, the decoder first performs a consensus step, effectively converting the N clustered output strings into $(1 - q_0)M$ strings with a smaller erasure rate. From here, the decoder proceeds similarly to the single-draw case. Each cluster is assigned a distinct label from $\{1, \ldots, M\}$, and the clusters are ordered by label to create a single output vector $\mathbf{y}$ of length roughly $(1 - q_0)ML$. All of the at most $M!$ possible label assignments are considered. The system of equations corresponding to each label assignment and vector $\mathbf{y}$ can be solved for a solution $\mathbf{t}$, if a solution exists. This cluster-based decoding scheme is illustrated in Figure 4.

Again, we must choose B large enough so that the true system, obtained by clustering and ordering the output strings correctly, has a unique solution. Additionally, we must have R small enough so that only one of the systems (the true one) yields a solution that corresponds with a valid message vector $\mathbf{t}_i$ so that the correct message can be decoded.

To guarantee a unique solution, the true system must have enough equations after erasures have been discarded. After the consensus step has been performed on the output clusters, we have at least $p_L M$ output strings and an expected effective erasure probability of p_{eff}. It can be shown using standard concentration inequalities that the effective erasure probability cannot deviate significantly from p_{eff}. Thus, the true system will have at least $MLp_L(1 - p_{\text{eff}} - \epsilon)$ equations with high probability, and if we set

$$B = MLp_L(1 - p_{\text{eff}} - \epsilon)(1 - \epsilon),$$

then by Lemma 1, the true system of equations will have a unique solution with probability tending to 1 as $M \to \infty$.

To find the maximum rate R for which this scheme succeeds with vanishing error probability, we must bound the total number of valid clusterings of the output strings. We begin by analyzing the total number of edges in the consistency graph. Since each cluster of size n produces $\binom{n}{2} \leq n^2/2$ "correct" edges (i.e., edges between output strings that originated from the same input string), the expected number of correct edges is at most

$$\sum_{n=0}^{\infty} M q_n \frac{n^2}{2} = \frac{M}{2} E[Q^2]. \tag{5}$$

Now let Z be the total number of incorrect edges in the consistency graph, and define

$$\gamma := -\beta \log\left(1 - \tfrac{1}{2}(1 - p)^2\right),$$

which is positive for any $p \in (0, 1)$.

Lemma 2. *The number of incorrect edges Z satisfies*

$$\Pr\left(Z > M^{2-\gamma+\epsilon}\right) \to 0 \tag{6}$$

as $M \to \infty$, for any $\epsilon > 0$.

From Lemma 2, we can see that as long as $\gamma > 1$, the number of incorrect edges grows slower than M, and will therefore be vanishingly small compared to the number of correct edges given in (5).

We now bound the number of possible matchings given the total number of edges and find that a loose bound is sufficient to establish capacity.

Lemma 3. *Suppose the consistency graph has a total of U edges. Then there are at most 2^U valid ways to cluster the output sequences.*

Proof. Each valid clustering corresponds to a partition of the output sequences such that each group corresponds to a

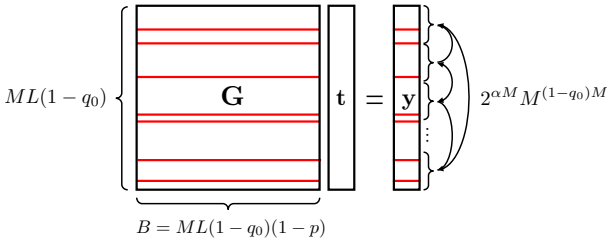

Fig. 5. In the multi-draw setting, there are $2^{\alpha M} M^{(1-q_0)M}$ potential systems of equations, obtained by considering all valid ways to cluster the output strings into $(1-q_0)M$ clusters, and then assigning each cluster to an index. The true system (corresponding to the correct clustering and ordering) has $ML(1-q_0)(1-p_{\text{eff}})$ non-erased equations in expectation.

clique in the consistency graph. Notice that a partition of the consistency graph into cliques is uniquely described by the set of edges that are part of each of the cliques. Hence, each partition corresponds to a distinct subset of the U edges in the graph. Since there are at most 2^U such subsets, it follows that there are at most 2^U valid ways to cluster the output sequences. □

Following (5) and Lemma 2, we know that for some $\gamma > 1$, the number of edges U in the consistency graph satisfies $U < \alpha M$ with high probability, for some $\alpha > 1$. Thus, Lemma 3 implies that the number of ways to cluster the output sequences is at most $2^{\alpha M}$.

Now, similar to the single-draw achievability proof (Section III-A), we must guarantee that there is no collision between the solution found $\mathbf{t}$ and any incorrect codeword $\mathbf{t}_i$, for $i = 2, \ldots, 2^{MLR}$ (assuming message 1 is sent). The decoder must attempt to solve a total of at most $2^{\alpha M} M^{p_U M}$ systems of equations, as we need to cluster the output strings into at most $p_U M$ valid clusters and assign each cluster an index in $\{1, \ldots, M\}$ for the ordering. The probability of a collision is then upper-bounded by

$$\begin{aligned} 2^{\alpha M} M^{p_U M} 2^{MLR} 2^{-B} &= 2^{\alpha M + p_U M \log M + MLR - B} \\ &= 2^{ML(\alpha/L + p_U/\beta + R - B/(ML))}, \end{aligned} \quad (7)$$

which goes to 0 as $M \to \infty$ as long as

$$R + \frac{\alpha}{\beta \log M} - p_L(1 - p_{\text{eff}} - \epsilon)(1-\epsilon) + \frac{p_U}{\beta} < 0.$$

Since $\alpha/(\beta \log M) \to 0$ and ϵ can be chosen arbitrarily small,

$$R < (1-q_0)(1 - p_{\text{eff}} - 1/\beta)$$

is achievable, and we have the following capacity lower bound:

Theorem 2. *The capacity of the BEC multi-draw DNA storage channel satisfies*

$$C_{\textit{BEC, multi-draw}} \geq (1-q_0)\left(1 - p_{\textit{eff}} - 1/\beta\right), \quad (8)$$

as long as $\gamma = -\beta \log\left(1 - \frac{1}{2}(1-p)^2\right) > 1$.

Since the parameter regime required for the upper bound in Equation 4 is smaller than that required for the lower bound, we have that Theorem 1 holds for the smaller regime $\beta > 2/(1 - 2p + p^2)$ (the blue region in Figure 2).

IV. Discussion

A natural follow-up to this work is to investigate whether the random linear coding scheme here presented also achieves the capacity of the BSC multi-draw DNA storage channel. This is reasonable, as linear codes are also known to achieve the capacity of a BSC [16]. However, this problem is more complicated than the case of the BEC, as the concept of consistency does not exist with bit substitution errors.

Note that the clustering algorithm covered here is "code-aware"; the decoder considers every feasible clustering of the output strings and only permits a clustering whose corresponding linear equations have a unique, valid solution (message index). Hence, another natural follow-up question is whether a code-oblivious clustering algorithm is also sufficient to achieve capacity. Code-oblivious clustering approaches are more desirable as they provide a separation between the clustering and decoding tasks. Notice that the greedy clustering algorithm used in [9] for the BSC case is code-oblivious.

Finally, we remark that the capacity for a large set of parameters (p, β) is still an open question. Weinberger and Merhav [10] provide the capacity for a general DMC, which in the BEC case corresponds to the blue region in Figure 2. Characterizing the rates achieved by linear coding schemes and code-aware clustering in the gray region in Figure 2 is another direction for future work.

Appendix A
Proof of Lemma 1

Lemma 1. *Let* $\mathbf{G}$ *be an* $ML \times B$ *matrix with i.i.d.* Bernoulli$(1/2)$ *entries. Fix any* $\delta \in (0,1)$ *and a submatrix* $\mathbf{G}'$ *formed by an arbitrary set of* $(1-\delta)B$ *rows of* $\mathbf{G}$. *Then* $\mathbf{G}'$ *is full rank (over the finite field* $\mathbb{F}_2$*) with probability tending to* 1 *as* $B \to \infty$.

Proof. We follow the approach in the lecture notes by [17]. In order for $\mathbf{G}'$ to be full rank, the $(n+1)$th row must be chosen as a vector that is not in the span of rows $1, \ldots, n$. Note that the space spanned by n linearly independent vectors in $\mathbb{F}_2$ has exactly 2^n distinct elements. If we assume that the first n rows are linearly independent, then the probability that the $(n+1)$th row (which is a B-dimensional vector) is not in the span of the first n rows is $1 - 2^{n-B}$. By induction we see that the probability that all $(1-\delta)B$ rows are linearly independent is

$$\begin{aligned} \prod_{j=1}^{(1-\delta)B} \left(1 - 2^{-(B-j+1)}\right) &= \prod_{i=\delta B+1}^{B} \left(1 - 2^{-i}\right) \\ &= \frac{\prod_{i=1}^{B}\left(1-2^{-i}\right)}{\prod_{i=1}^{\delta B}\left(1-2^{-i}\right)}. \end{aligned} \quad (9)$$

As $B \to \infty$, both the product in the numerator and in the denominator can be verified (e.g., using numerical software) to converge to

$$\prod_{i=1}^{\infty}\left(1-2^{-i}\right) = 0.28879...$$

which implies that (9) tends to 1 as $B \to \infty$, proving the lemma. Notice that, if $\delta = 0$, the probability does not tend to 1 and instead tends to 0.28879, which is in contrast to the case of a real-valued matrix with random entries from a continuous distribution, where all square submatrices will be full-rank with high probability. □

APPENDIX B
PROOF OF LEMMA 2

Lemma 2. *The number of incorrect edges Z satisfies*

$$\Pr\left(Z > M^{2-\gamma+\epsilon}\right) \to 0 \tag{10}$$

as $M \to \infty$, for any $\epsilon > 0$.

Proof. Consider two output strings y_i^L and y_j^L that are generated from distinct input strings x_i^L and x_j^L. We show that y_i^L and y_j^L are consistent with probability $M^{-\gamma}$.

Let $x_i^L[\ell]$ and $x_j^L[\ell]$, for $\ell = 1, \ldots, L$ be the individual symbols in the sequences. Notice that x_i^L and x_j^L are generated by choosing $\mathbf{t}_i$ and $\mathbf{t}_j$ uniformly at random from $\{0,1\}^B$, and computing $\mathbf{G}'\,\mathbf{t}_i$ and $\mathbf{G}''\,\mathbf{t}_j$, where $\mathbf{G}'$ and $\mathbf{G}''$ are each obtained by taking the L rows from $\mathbf{G}$ corresponding to the ith and jth input sequences (note that $\mathbf{G}\mathbf{t}_i$ has length ML).

First we claim that the $2L$ random variables $x_i^L[\ell], x_j^L[\ell]$, $\ell = 1, \ldots, L$, are mutually independent Bernoulli$(1/2)$. Treating all vectors as column vectors, we can write

$$\begin{bmatrix} x_i^L \\ x_j^L \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{G}' & 0 \\ 0 & \mathbf{G}'' \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} \mathbf{t}_i \\ \mathbf{t}_j \end{bmatrix}}_{\tilde{\mathbf{t}}}. \tag{11}$$

The block diagonal matrix H above has dimension $2L \times 2B$, where $B = ML(1-q_0-\epsilon)(1-p_{\text{eff}}-\epsilon)(1-\epsilon)$. For M large enough, we have $B > L$, and H is full row-rank, with a null space of dimension $2B - 2L$. Hence, for any $\mathbf{c} \in \mathbb{F}_2^L$, the number of solutions $\tilde{\mathbf{t}}$ to $\mathbf{c} = H\tilde{\mathbf{t}}$ is 2^{2B-2L} and, if $\tilde{\mathbf{t}}$ is drawn uniformly at random from $\mathbb{F}_2^{2B}$,

$$\Pr(H\tilde{\mathbf{t}} = \mathbf{c}) = \frac{2^{2B-2L}}{2^{2B}} = 2^{-2L}.$$

This implies that the column vector $\begin{bmatrix} x_i^L \\ x_j^L \end{bmatrix}$ is chosen uniformly at random from $\mathbb{F}_2^{2L}$. This in turn implies that the entries of x_i^L and x_j^L are all mutually independent i.i.d. Bernoulli$(1/2)$ random variables.

Given this fact, the event that y_i^L and y_j^L are consistent is the intersection of L independent events

$$\{x_i^L[\ell] = x_j^L[\ell] \text{ or } x_i^L[\ell] = \varepsilon \text{ or } x_j^L[\ell] = \varepsilon\},$$

for $\ell = 1, \ldots, L$. Each of these events happens with probability $1 - \frac{1}{2}(1-p)^2$, implying that x_i^L and x_j^L are consistent with probability

$$(1 - \tfrac{1}{2}(1-p)^2)^L = 2^{-\gamma \log M} = M^{-\gamma}.$$

Finally, we notice that the expected number of output sequences is $ME[N_1]$, and the expected number of pairs of output strings is at most $M^2E[N_1]^2$. Hence, the expected number of incorrect edges satisfies

$$E[Z] \le M^2E[N_1]^2M^{-\gamma} = E[N_1]^2M^{2-\gamma}.$$

Finally, using Markov's inequality, we have that

$$\begin{aligned}\Pr\left(Z > M^{2-\gamma+\epsilon}\right) &\le \frac{E[Z]}{M^{2-\gamma+\epsilon}} \le \frac{E[N_1]^2M^{2-\gamma}}{M^{2-\gamma+\epsilon}} \\ &= E[N_1]^2M^{-\epsilon},\end{aligned}$$

which tends to 0 as $M \to \infty$ for any $\epsilon > 0$. □

REFERENCES

[1] G. M. Church, Y. Gao, and S. Kosuri, "Next-Generation Digital Information Storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628, 2012.

[2] N. Goldman, P. Bertone, and S. et al Chen, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, Feb. 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," *Angewandte Chemie International Edition*, vol. 54, no. 8, Feb. 2015.

[4] H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, , and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Sci. Rep.*, vol. 5, pp. 1–10, Sep. 2015.

[5] Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, Mar. 2017.

[6] 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 Biotechnol.*, vol. 36, no. 3, pp. 242–248, 2018.

[7] P. L. Antkowiak et al., "Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction," *Nature Commun.*, vol. 11, no. 1, pp. 1–10, Dec. 2020.

[8] 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.

[9] 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, pp. 8846–8850.

[10] N. Weinberger and N. Merhav, "The DNA Storage Channel: Capacity and Error Probability," 2021.

[11] I. Shomorony and R. Heckel, "DNA-Based Storage: Models and Fundamental Limits," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3675–3689, 2021.

[12] R. Heckel, I. Shomorony, K. Ramchandran, and N.C. David, "Fundamental limits of DNA storage systems," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 3130–3134.

[13] I. Shomorony and R. Heckel, "Capacity results for the noisy shuffling channel," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 762–766.

[14] S. Shin, R. Heckel, and I. Shomorony, "Capacity of the Erasure Shuffling Channel," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8841–8845.

[15] T. Shinkar, E. Yaakobi, A. Lenz, and A. Wachter-Zeh, "Clustering-Correcting Codes," 2019.

[16] P. Elias, "Coding for two noisy channels," in *Information Theory, 3rd London Symposium*, London, England, 1955.

[17] J. Sayir, "Lecture Notes for Advanced Communications and Coding: Binary Linear Codes over the Erasure Channel," 2014.