On the Locality in Codes for DNA Storage

Siyi Yang[1], *Student Member, IEEE*, Clayton Schoeny[1], *Student Member, IEEE*,
Laura Conde-Canencia[2], and Lara Dolecek[1], *Senior Member, IEEE*
[1] Department of Electrical and Computer Engineering, University of California, Los Angeles, Los Angeles, CA 90095 USA
[2] Lab-STICC, CNRS UMR 6285, Université de Bretagne Sud, Lorient, France

***Abstract*—Recently, error-correcting codes for DNA storage have been intensely studied. In DNA storage, information is stored as a prescribed number of pairwise distinct DNA molecules, each of length L. Lenz *et al.* showed that codes that correct up to s losses of DNA strands and t edited strands have a redundancy of at least $(s+t)L$ symbols. Based on the framework of Lenz *et al.*, we first present an explicit construction of codes with redundancy $O((s+2t)L)$ symbols. Locality is important for a rewritable random-access DNA storage system that tolerates frequent updates and moderate edits. By locality, we refer to the capability of a code to decode the original message without requiring the entire set of coded DNA strands. With this goal in mind, we then extend our previous code into a locally recoverable construction. Next, we focus on linear block codes that offer good trade-off between local distance and global distance. In this context, local distance refers to the minimum Hamming distance of each block and global distance refers to that of a prescribed number of consecutive blocks. Lastly, for a given local minimum distance and redundancy, we prove the existence of codes that reach the upper bound on the global minimum distance.**

I. Introduction

DNA storage systems have garnered substantial research interest recently because of their potential to store large amounts of data. It has been shown by Zhirnov *et al.* that increasing demand will exceed the available supply of silicon-based memories in the future, which motivates the exploration for new storage mediums [1].

There are two major operations in DNA storage systems: DNA synthesis and DNA sequencing, which correspond to writing and reading, respectively. Through DNA synthesis, binary files are encoded and stored as short strands of nucleotides. This information can be accessed through DNA sequencing, where multiple duplications of the nucleotides are generated through the polymerase chain reaction (PCR) technique. The original data can then be decoded by the obtained set of replicas. Recent progress on DNA synthesis and sequencing has bridged the gap between theory and practical implementations. As a consequence, research groups from Microsoft Research, Harvard University, and Washington University have reported successful implementation of DNA storage systems [2]–[5].

Although DNA storage systems are expected to have long-term stability, research on DNA synthesis and sequencing indicates that stored data suffers from various types of errors [4], [5]; information is lost from unsuccessful synthesis and hydrolytic damage in storage [6]. Synthesizing and sequencing DNA may also lead to insertions, deletions and substitutions of nucleotides. These error patterns motivate us to develop novel coding schemes that can appropriately correct the errors in DNA storage systems.

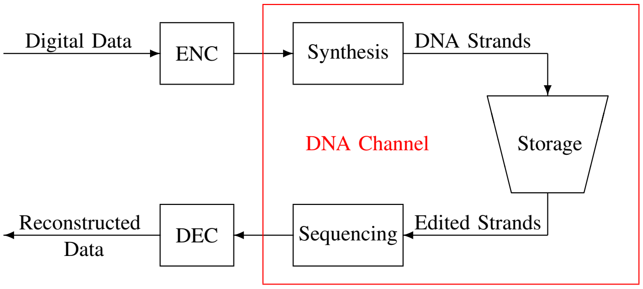

Fig. 1. DNA Storage System

Error-correcting codes in DNA storage systems have been intensely studied during recent years. Gabrys *et al.* studied codes for DNA storage systems in the asymmetric Lee distance and the Damerau distance [7], [8]. More recently, Lenz *et al.* introduced the concept of coding over sets and proposed constructions that are close to the optimal rate [9].

A major differentiator between DNA storage and conventional storage is that DNA strands are stored without the ordering information. The scheme in [9] innovatively interprets sets of strands as codewords, resulting in a rate close to optimal. However, this scheme requires decoding the whole file, even if only a small specific section needs to be read. Moreover, the scheme is not distance preserving, namely, rewriting even a few bits requires editing the vast majority of stored nucleotides. These properties prevent DNA storage from being used in frequently updating systems that require moderate editing. Yazdi *et al.* and Organick *et al.* studied rewritable, random-access DNA storage systems [3], [10], where a set of mutually uncorrelated strings (also called primers) were attached to different information blocks to enable access of information at arbitrary positions. Later on, Levy and Yaakobi proposed efficient algorithms for constructing mutually uncorrelated codes that can be appropriately used as primers [11]. Jain *et al.*, and Chee *et al.* studied coding schemes that correct tandem repeats, which addressed the errors that occur in PCR amplification resulting from the secondary structure in primers [12], [13].

In this paper, we first introduce the system model in Section II. Next, in Section III, we propose an explicit construction of codes for DNA storage that have a redundancy that scales linearly to the optimal redundancy, based on the study of Lenz

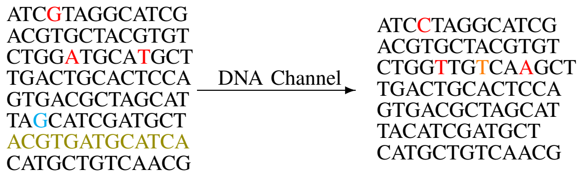

Fig. 2. DNA Channel: DNA suffers from errors including substitutions, insertions and deletions, as well as sequence losses, which are marked as red, orange, blue, and olive, respectively.

et al. [9]. This construction has an order-optimal rate, but no favorable locality properties. We then extend this construction into one that is locally recoverable. In Section IV, we study constructions of locally recoverable linear block codes defined on a finite field $GF(q)$. We prove the existence of linear block codes that offer a good trade-off between the local Hamming distance and the global Hamming distance, for sufficiently large field size q. Finally, we summarize and discuss future extensions. The $\log$ base is always 2 throughout this paper.

II. System Model

A. DNA Storage System

DNA storage systems involve two stages: reading and writing. In the writing stage, binary data is encoded into a set of short strings that consist of approximately $150 \sim 250$ nucleobases chosen from $\{A, T, C, G\}$. As shown in Fig. 1, the *de novo* DNA systhesis technique is then applied to synthesize the artificially designed strands into real nucleotides [14]. In the reading stage, the nucleotides are sampled and read by a DNA sequencing technique. Next Generation Sequencing (NGS), including Illumina sequencing and Nanopore sequencing techniques, allows for massive parallel sequencing of nucleotides at a high throughput and a low error probability. The original information is then decoded from the content of the sequenced nucleotides. Throughout this paper, we focus on DNA storage based on Illumina sequencing [6].

In Fig. 1, we present a block diagram of the DNA channel, highlighting the synthesis, storage and sequencing steps. Fig. 2 shows the input, output and the error model of the DNA channel. Current synthesis techniques generate multiple duplications of the targeted strands, implying that information stored in identical strands cannot be distinguished during the reading steps [15]. Additionally, during storage, DNA strands undergo breaking and bridge amplification due to depurination in storage caused by hydrolytic damage. These DNA molecules cannot be read by Illumina sequencing, resulting in a significant loss of DNA strands at the output of the DNA channel [6]. Moreover, NGS sequencers also introduce substitutions, insertions and deletions within nucleotides [14], [16].

B. Characterization of the DNA Channel

According to the DNA channel described in [6], [9], the input and output of a DNA channel can be modeled as follows.

Let $\mathcal{X}_M^L$ be the set of all subsets consisting of M strings of L symbols from an alphabet of size 4. In DNA storage, particularly, the alphabet is $\{A, T, C, G\}^L$. Then the input, i.e., any set S consisting of M strands of length L, must be an element of $\mathcal{X}_M^L$. Note that identical strands cannot be distinguished by DNA sequencing since the strands are stored without ordering information and similar strands are clustered together as edited replicas of a single strand.

The output of the DNA channel is a set S' obtained from S through a loss of at most s strands and edits in at most t strands. In Fig. 2, for example, $M = 8$, $L = 13$, $s = 1$, and $t = 3$.

III. Constructions

For the remainder of the paper, we use the DNA channel model introduced in Section II-B. Lenz *et al.* analyzed the optimal rate of codes that correct entire losses of DNA strands as well as edits within DNA strands. They show how constant weight codes that correct asymmetric errors can be used as constituent codes in a coding scheme for the DNA channel. This construction is not explicit and is of high complexity. In this section, we provide an explicit construction of a code that has redundancy up to a constant factor times the optimal redundancy. This construction can also be extended to a DNA storage system that is locally recoverable.

A. Rate-optimal Codes

Definition 1. (cf. [9]) *A code $\mathcal{C} \subset \mathcal{X}_M^L$ is called an (s,t) error-correcting code, if it can correct up to s losses of sequences and edits within t sequences.*

Definition 2. *For any integers $N, l, a, m \in \mathbb{N}^*$, where $l \leq N$, $a < l$, the set $A(N, l, a) = \{A_1, A_2, \cdots, A_m\}$ of m subsets of $[0 : N-1]$ is called an **(N, l, a)-set** if for all $1 \leq i \leq m$,*

1) *$|A_i| = l$.*
2) *$\forall j \neq i$ and $j \in [M]$, $|A_i \setminus A_j| > a$.*

*A code $\mathcal{C} \subset \mathcal{X}_M^L$ is called an (M, L, d) **code** if it is a $(4^L, M, d)$-set.*

We know from [9] that any $(M, L, s+2t)$ code is also an (s,t) error-correcting code. Therefore our objective is to find efficient (M, L, d)-error-correcting codes.

Lemma 1. (cf. [9]) *The optimal redundancy $r(\mathcal{C})$ for any (s,t) code $\mathcal{C} \subset \mathcal{X}_M^L$ satisfies*

$$r(\mathcal{C}) \geq (s+t)\log(2^L - M - t) + t\log(M - s - t) - \log(t!(s+t)!).$$

Lemma 1 indicates that the optimal rate of an (s,t)-code has redundancy of $O((s+t)L)$ symbols. Since each strand has length L, this means that $O(s+t)$ out of M strands carry the redundant information in the optimal case. In Construction 1, we provide an (M, L, d) code with $O(d)$ redundant strands, which is an (s,t) code with $O(s+2t)$ redundant strands when $d = s + 2t$.

Prior to Construction 1, we require Lemma 2 (from our previous work [17]), in which we define a function $\alpha^{(q,d)}$, called (q,d)**-parity**, that maps the subsets of $GF(q)$ onto $GF(q)^{2d-1}$. It is shown that the cardinality of the set difference between any two distinct subsets of $GF(q)$ of the same size is greater than d if their (q,d)-parities are identical.

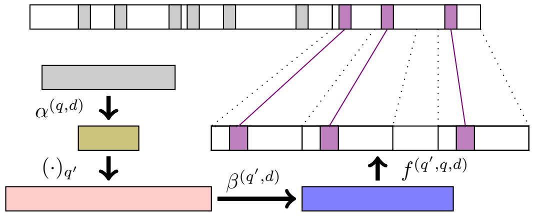

Fig. 3. Diagram of Construction 1.

Lemma 2. (cf. [17]) *Define the* (q,d)*-parity of any set* $S \subset GF(q)$ *as follows,*

$$\begin{cases} \alpha_1 &= \sum_{s\in S} s, \\ \alpha_2 &= \sum_{s\in S} s^2, \\ &\vdots \\ \alpha_{2d-1} &= \sum_{s\in S} s^{2d-1}. \end{cases} \tag{1}$$

For all $A, B \subset GF(q)$, $|A| = |B|$, *if* $\alpha^{(q,d)}(A) = \alpha^{(q,d)}(B)$, *then* $|A \setminus B| > d$.

Fig. 3 depicts the main idea of Construction 1. We construct a code $\mathcal{C}_1$ with the assistance of an auxiliary code $\mathcal{C}_H(n,(2k-1)d,d+1)_{q'}$, which is defined on a smaller field $GF(q')$ (the auxiliary field), $q' \le q^{\frac{1}{2}}$, $q' = 2^{L'}$ for some $L' \in \mathbb{N}$, has minimum Hamming distance $d+1$, and encodes $(2k-1)d$ symbols into n symbols, where $k = \lceil \frac{2L}{\log q'} \rceil$. Here, we partition the nonzero elements of $GF(q)$, $q = 4^L$, into two parts I_1, I_2. Among the M strands contained in $S \subset \mathcal{X}_M^L$, $M-n$ strands from I_1 carry the raw information, and the remaining n strands from I_2 carry the redundant information.

1) For any subset A of I_1 with cardinality $M-n$, compute the (q,d)-parity $\alpha^{(q,d)}(A)$. $\alpha^{(q,d)}(A)$ is a vector on $GF(q)$ of length $2d-1$.
2) Represent $\alpha^{(q,d)}(A)$ as a vector on $GF(q')$ of length $k(2d-1)$ and denote it by $\left(\alpha^{(q,d)}(A)\right)_{q'}$.
3) $\beta^{(q',d)}$ maps $\left(\alpha^{(q,d)}(A)\right)_{q'}$ into a codeword $\boldsymbol{c}(A)$ in $\mathcal{C}_H(n,(2k-1)d,d+1)_{q'}$. $\boldsymbol{c}(A)$ is a vector of length n on $GF(q')$.
4) $f^{(q',q,d)}$ maps $\boldsymbol{c}(A)$ to a subset $f(A)$ of I_2 with cardinality n.
5) The codeword of A is $g(A) = A \cup f(A)$.

In Lemma 3, we provide an injection $f^{(q',q,d)}$ so that the image of $f^{(q',q,d)}$ is an (nq',n,d)-set. Then for any distinct $A, B \subset I_1$, either $|A \setminus B| > d$ or $|f(A) \setminus f(B)| > d$, which means that the resultant $|g(A) \setminus g(B)| > d$. Let $d = s+2t$, we construct an (s,t) code with $n = k(2d-1)+d = (2k+1)(s+2t)-k$ redundant strands. When $k=2$ and $q' = 2^L$, there are $5(s+2t)-2$ redundant strands, and the code is order-optimal.

Lemma 3. *Suppose* M, L, d *are given,* $q = 4^L$. *Given* $q' \le q^{\frac{1}{2}}$, *let* $k = \lceil \frac{2L}{\log q'} \rceil$, $n = (2k+1)d-k$. *Suppose* $\mathcal{C}_H(n,n-d,d+1)_{q'}$ *is a Reed Solomon code of length* n *with* d *redundant symbols. There is a bijection* $\gamma : [n] \times GF(q') \to [nq']$. *Define* $f^{(q',q,d)} : \mathcal{C}_H(n,n-d,d+1)_{q'} \to [nq']$ *by* $f^{(q',q,d)} : (c_1, c_2, \cdots, c_n) \to \{\gamma((i,c_i)) | 1 \le i \le n\}$. *Then* $Im(f)$, *i.e., the image of* f, *is an* (nq',n,d)*-set.*

Proof. For any distinct $\boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathcal{C}_H(n,n-d,d+1)_{q'}$, we only need to prove $|f^{(q',q,d)}(\boldsymbol{c}_1) \setminus f^{(q',q,d)}(\boldsymbol{c}_2)| \ge d+1$. Since $\mathcal{C}_H(n,n-d,d+1)_{q'}$ is a Reed Solomon code, $\boldsymbol{c}_1 \ne \boldsymbol{c}_2$, $d_H(\boldsymbol{c}_1,\boldsymbol{c}_2) \ge d+1$. Suppose $\boldsymbol{c}_1 = (c_{1,1}, c_{1,2}, \cdots, c_{1,n})$, $\boldsymbol{c}_2 = (c_{2,1}, c_{2,2}, \cdots, c_{2,n})$. Since γ is an injection, $|f^{(q',q,d)}(\boldsymbol{c}_1) \setminus f^{(q',q,d)}(\boldsymbol{c}_2)| = |\{(i,c_{1,i}) | 1 \le i \le n\} \setminus \{(i,c_{2,i}) | 1 \le i \le n\}| = |\{i | c_{1,i} \ne c_{2,i}, 1 \le i \le n\}| = d_H(\boldsymbol{c}_1,\boldsymbol{c}_2) \ge d+1$. ■

We denote the (M,L,d)-error correcting codes constructed with an auxiliary field $GF(q')$ by $\mathcal{C}_1(M,L,d)_{q'}$.

Construction 1. *Suppose* M, L, d *are given, and let* $q = 4^L$. *Given* $q' \le q^{\frac{1}{2}}$, *let* $k = \lceil \frac{2L}{\log q'} \rceil$, $n = (2k+1)d-k$. *Define* g *on the set of all subsets of* $\left[nq' : 4^L - 1\right]$ *by* $g : A \mapsto A \cup \left(f^{(q',q,d)}\left(\boldsymbol{c}(A)\right)\right)$, *where* $\boldsymbol{c}(A)$ *is the codeword* $\beta^{(q',k,d)}\left(\left(\alpha^{(q,d)}(A)\right)_{q'}\right)$. *Then* $\mathcal{C}_1(M,L,d)_{q'} = \{g(A) | A \subset \left[nq', 4^L - 1\right], |A| = M-n\}$ *is an* (M,L,d)*-code.*

Proof. For any $A, B \subset \left[nq', 4^L - 1\right]$, $|A| = |B| = M-n$, $A \ne B$, we only need to prove that $|g(A) \setminus g(B)| \ge d+1$. For simplicity, we denote $f^{(q',q,d)}$ by f in this proof. Notice that neither $f(\boldsymbol{c}(A))$ nor $f(\boldsymbol{c}(B))$ has nonzero intersection with either A or B. Therefore $|g(A) \setminus g(B)| = |A \setminus B| + |f(\boldsymbol{c}(A)) \setminus f(\boldsymbol{c}(B))|$. There are two cases:

1) $\alpha^{(q,d)}(A) = \alpha^{(q,d)}(B)$. Then Lemma 2 implies that $|A \setminus B| \ge d+1$. Then $|g(A) \setminus g(B)| \ge |A \setminus B| \ge d+1$.
2) $\alpha^{(q,d)}(A) \ne \alpha^{(q,d)}(B)$. Then the q'-ary representations $\left(\alpha^{(q,d)}(A)\right)_{q'} \ne \left(\alpha^{(q,d)}(B)\right)_{q'}$, and their corresponding codewords $\boldsymbol{c}(A), \boldsymbol{c}(B)$ in $\mathcal{C}_H((2k+1)d-k, k(2d-1), d+1)_{q'}$ are non-equal. Then Lemma 3 implies that $|f(\boldsymbol{c}(A)) \setminus f(\boldsymbol{c}(B))| > d$, thus $|g(A) \setminus g(B)| \ge |f(\boldsymbol{c}(A)) \setminus f(\boldsymbol{c}(B))| \ge d+1$.

■

Example 1. *Let* $M = 60$, $L = 4$, $d = 3$. *Then* $q = 4^L = 256$. *Let* $q' = 16$. *Therefore* $k = 2$, $n = (2k+1)d - k = 13$, $|A| = M - n = 47$, $A \subset \left[nq' : q-1\right] = [208 : 255]$. *Then the number of codewords is* $\binom{255-208+1}{47} = 48$. *Suppose* $A(s) = [208 : 255] \setminus \{207 + s\}$, $1 \le s \le 48$. *Then the corresponding codeword of* $A(s)$ *is* $c_s = A(s) \cup f(A(s))$, $1 \le s \le 48$, *where* $f(A(s))$ *are defined as follows, using Construction 1:*

$$\begin{aligned}
f(A(1)) &= \{16, 32, 47, 51, 67, 87, 107, 127, 135, 152, 166, 181, 203\}, \\
f(A(2)) &= \{15, 17, 43, 59, 68, 85, 99, 116, 134, 159, 172, 189, 206\}, \\
f(A(3)) &= \{15, 18, 43, 60, 67, 87, 99, 115, 134, 149, 168, 183, 195\}, \\
f(A(4)) &= \{15, 19, 43, 63, 77, 81, 100, 116, 134, 149, 162, 184, 201\}, \\
f(A(5)) &= \{15, 20, 43, 64, 78, 85, 100, 115, 133, 157, 163, 184, 199\}, \\
f(A(6)) &= \{15, 21, 44, 59, 78, 95, 100, 127, 138, 154, 167, 187, 198\}, \\
f(A(7)) &= \{15, 22, 44, 60, 78, 89, 100, 128, 137, 155, 170, 191, 194\}, \\
f(A(8)) &= \{15, 23, 44, 63, 66, 91, 99, 127, 143, 154, 162, 185, 202\}, \\
f(A(9)) &= \{15, 24, 44, 64, 66, 91, 99, 128, 143, 153, 162, 183, 205\},
\end{aligned}$$

$$\begin{aligned}
f(A(10)) &= \{15,25,47,59,71,87,111,127,142,150,168,188,205\},\\
f(A(11)) &= \{15,26,47,60,68,93,111,128,130,155,166,188,193\},\\
f(A(12)) &= \{15,27,47,63,68,83,112,127,142,153,169,192,203\},\\
f(A(13)) &= \{15,28,47,64,71,95,112,128,129,150,166,182,200\},\\
f(A(14)) &= \{15,29,48,59,66,82,112,116,143,154,169,181,198\},\\
f(A(15)) &= \{15,30,48,60,70,96,112,115,132,160,170,187,195\},\\
f(A(16)) &= \{15,31,48,63,72,86,111,116,138,159,171,186,203\},\\
f(A(17)) &= \{15,32,48,64,68,94,111,115,134,155,165,190,205\},\\
f(A(18)) &= \{14,17,45,52,74,94,110,124,134,147,165,177,199\},\\
f(A(19)) &= \{14,18,45,51,78,87,110,123,138,145,172,184,202\},\\
f(A(20)) &= \{14,19,45,56,72,92,109,124,138,147,163,189,193\},\\
f(A(21)) &= \{14,20,45,55,68,87,109,123,133,147,163,178,207\},\\
f(A(22)) &= \{14,21,46,52,79,90,109,119,130,153,172,190,202\},\\
f(A(23)) &= \{14,22,46,51,76,87,109,120,141,148,168,181,206\},\\
f(A(24)) &= \{14,23,46,56,68,96,110,119,139,147,161,185,199\},\\
f(A(25)) &= \{14,24,46,55,71,87,110,120,135,156,164,188,196\},\\
f(A(26)) &= \{14,25,41,52,68,93,98,119,138,159,169,190,202\},\\
f(A(27)) &= \{14,26,41,51,68,96,98,120,138,154,170,177,198\},\\
f(A(28)) &= \{14,27,41,56,72,91,97,119,134,154,172,191,205\},\\
f(A(29)) &= \{14,28,41,55,72,96,97,120,133,157,166,186,194\},\\
f(A(30)) &= \{14,29,42,52,78,86,97,124,131,160,166,186,200\},\\
f(A(31)) &= \{14,30,42,51,77,83,97,123,132,146,168,187,193\},\\
f(A(32)) &= \{14,31,42,56,75,84,98,124,138,147,172,180,204\},\\
f(A(33)) &= \{14,32,42,55,76,83,98,123,138,159,167,187,206\},\\
f(A(34)) &= \{13,17,46,63,74,81,106,120,140,155,175,186,203\},\\
f(A(35)) &= \{13,18,46,64,78,87,106,119,131,149,166,179,207\},\\
f(A(36)) &= \{13,19,46,59,65,93,105,120,144,160,166,181,193\},\\
f(A(37)) &= \{13,20,46,60,69,93,105,119,136,148,162,182,198\},\\
f(A(38)) &= \{13,21,45,63,73,94,105,123,130,145,174,181,193\},\\
f(A(39)) &= \{13,22,45,64,78,96,105,124,138,152,166,178,208\},\\
f(A(40)) &= \{13,23,45,59,67,82,106,123,131,160,172,177,196\},\\
f(A(41)) &= \{13,24,45,60,72,86,106,124,140,155,173,192,206\},\\
f(A(42)) &= \{13,25,42,63,75,94,102,123,135,145,170,186,199\},\\
f(A(43)) &= \{13,26,42,64,75,84,102,124,132,156,173,185,194\},\\
f(A(44)) &= \{13,27,42,59,74,82,101,123,131,147,168,188,208\},\\
f(A(45)) &= \{13,28,42,60,74,90,101,124,135,156,174,177,202\},\\
f(A(46)) &= \{13,29,41,63,67,94,101,120,132,146,169,190,206\},\\
f(A(47)) &= \{13,30,41,64,68,88,101,119,136,148,175,179,194\},\\
f(A(48)) &= \{13,31,41,59,67,82,102,120,129,154,172,183,206\}.
\end{aligned}$$

In this example, we have $|c_i \setminus c_j| \geq d+1 = 4$, for all $1 \leq i < j \leq 48$.

In Theorem 1, we analyze the redundancy of $\mathcal{C}_1(M,L,d)_{q'}$. Compared to the lower bound provided by Lemma 1, the redundancy of our construction is up to a constant factor times the smallest redundancy.

Theorem 1. *Given $q' \leq q^{\frac{1}{2}}$, let $k = \lceil \frac{2L}{\log q'} \rceil$, $n = (2k+1)d - k$. If $Mq' < 4^L - M + n$, then $\mathcal{C}_1(M,L,d)_{q'}$ has redundancy $r(\mathcal{C}_1(M,L,d)_{q'}) < n(2L+1)$ bits.*

Proof. The redundancy of $\mathcal{C}_1(M,L,d)_{q'}$ can be computed by the following equation:

$$\begin{aligned}
r(\mathcal{C}_1(M,L,d)_{q'}) &= \log \binom{4^L}{M} - \log \binom{4^L - nq'}{M-n} \\
&= \sum_{i=1}^{M} \log(4^L - i + 1) - \sum_{i=1}^{M-n} \log(4^L - nq' - i + 1) \\
&\quad - \log M! + \log(M-n)! \\
&= \sum_{i=1}^{M-n} \log\left(1 + \frac{nq'}{4^L - nq' - i + 1}\right)
\end{aligned}$$

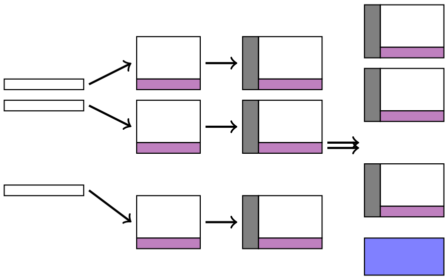

Fig. 4. Structure of Locally Recoverable Codes for DNA Storage.

$$
\begin{aligned}
&+\sum_{i=0}^{n-1}\log(4^L-M+n-i)-\sum_{i=1}^{n}\log(M-i+1)\\
<&\frac{nq'(M-n)}{4^L-nq'-M+n}+2nL\\
<&n+2nL=n(2L+1).
\end{aligned}
$$

■

Theorem 1 indicates that $\mathcal{C}_1(M,L,d)_{q'}$ has an order-optimal redundancy. This code is explicitly constructible. In the decoding process, we first compute the (q,d)-parity $\alpha^{(q,d)}(A)$, and then we derive the set A by the algorithm in [17].

B. Locally Recoverable Codes

Although Construction 1 has a high code-rate, both the decoding and writing procedures involve decoding and writing the entirety of the data stored in the system, which is not efficient when only a particular part of the data is of interest. In Construction 2, we provide a code $\mathcal{C}_2$ where the information is stored in m disjoint subsets of a codeword, and each subset has minimum set difference cardinality d_1+1 when $\mathcal{C}_2$ has minimum set difference cardinality d_2+1. The rate of the code is analyzed in Theorem 2.

Definition 3. *For any code $\mathcal{C}\subset\mathcal{X}_M^L$, $d_1,d_2\in\mathbb{N}$, $d_1<d_2$: suppose $D=\{\boldsymbol{p}_i\}_{i=1}^m$ is a set of distinct strings of symbols from an alphabet of size 4 with length t. Then, $\mathcal{C}$ is called an (m,M,L,d_1,d_2)-**error-correcting code** if the following conditions hold:*

1) *$\mathcal{C}$ itself is an (M,L,d_2)-code;*
2) *Each A_i is an $(|A_i|,L_1,d_1)$-code, where $A_i=\{T|T=\{\boldsymbol{c}|(\boldsymbol{p}_i,\boldsymbol{c})\in S,S\in\mathcal{C}\}\}$, $L_1=L-t$, for $1\le i\le m$.*

Fig. 4 depicts the main idea of Construction 2, where a construction of an (m,M,L,d_1,d_2)-code is presented. The original file is divided into m different parts, where each part is encoded into a set S_i of $\frac{M-l_2}{m}$ strands by the code $\mathcal{C}_1(\frac{M-l_2}{m},L-t,d_1)_{q_1}$, where $\frac{M-l_2}{m}$ is an integer. Then each strand in S_i is appended to a primer $\boldsymbol{p}_i$ of length t, for all $1\le i\le m$, and we obtain a set S of $M-l_2$ distinct strands of length L, from which we can obtain a set of M distinct strands of length L by the code $\mathcal{C}_1(M,L,d_2)_{q_2}$. Here q_1, q_2 refer to the size of the auxiliary field of $\mathcal{C}_1(\frac{M-l_2}{m},L-t,d_1)_{q_1}$ and $\mathcal{C}_1(M,L,d_2)_{q_2}$, respectively. We denote this code by $\mathcal{C}_2(m,M,L,d_1,d_2)_{q_1,q_2}$, and $\mathcal{C}_2$ is defined on $GF(q)$. In order to decode the i-th part, we only need to filter out the strands with primer $\boldsymbol{p}_i$, which can be done with current sequencing methods [3]. Then the information can be decoded since the remaining parts of those strands constitute an $(\frac{M-l_2}{m},L-t,d_1)$-code. When the local decoding fails, we can still decode by accessing all the strands, since the overall code is an (M,L,d_2) code.

Construction 2. *Suppose $m,d_1,d_2,M,L,q_1,q_2\in\mathbb{N}$ are given, $q=4^L$, and $q_1,q_2<q^{\frac{1}{2}}$. Find $t\in\mathbb{N}$ such that $t\ge\frac{1}{2}\log m$. Let $L_1=L-t$, $k_1=\lceil\frac{2L_1}{\log q_1}\rceil$, $l_1=(2k_1+1)d_1-k_1$, $k_2=\lceil\frac{2L}{\log q_2}\rceil$, $l_2=(2k_2+1)d_2-k_2$. Suppose $m|(M-l_2)$. Suppose D is a set of m distinct strings of symbols from an alphabet of size 4 with length t. Let $E=\{S|S=\{(\boldsymbol{p},\boldsymbol{c})|\boldsymbol{p}\in D,\boldsymbol{c}\in\mathcal{C}_1(q_1,\frac{M-l_2}{m},L_1,d_1)\}\}$ and $\mathcal{C}_2(m,M,L,d_1,d_2)_{q_1,q_2}=\{S\cup f^{(q_2,q,d_2)}(S)|S\in E\}$. Then $\mathcal{C}_2(m,M,L,d_1,d_2)_{q_1,q_2}$ is an (m,M,L,d_1,d_2) code.*

Theorem 2. *If $Mq_2<4^L-M+l_2$, $(\frac{M-l_2}{m})q_1<4^{L_1}-\frac{M-l_2}{m}+l_1$, and $q_1\ge 4$, then $\mathcal{C}_2(m,M,L,d_1,d_2)_{q_1,q_2}$ has redundancy $r(\mathcal{C}_2(m,M,L,d_1,d_2)_{q_1,q_2})<l_2(2L+1)+ml_1(2L_1+1)+M(2t-\log m+1)$ bits.*

Proof. Let $k_1=\lceil\frac{2L_1}{\log q_1}\rceil$, $k_2=\lceil\frac{2L}{\log q_2}\rceil$, $l_1=(2k_1+1)d_1-k_1$, $l_2=(2k_2+1)d_2-k_2$. We know that $\frac{M-l_2}{m}\ge l_1+1\ge 5d_1-2+1\ge 4$. The redundancy of $\mathcal{C}_2(m,M,L,d_1,d_2)_{q_1,q_2}$ can be computed by the following equation:

$$
\begin{aligned}
&r(\mathcal{C}_2(m,M,L,d_1,d_2)_{q_1,q_2})\\
=&\log\binom{4^L}{M}-m\log\binom{4^{L_1}-l_1q_1}{\frac{M-l_2}{m}-l_1}\\
=&\log\binom{4^L}{M}-\log\binom{4^L-l_2q_2}{M-l_2}\\
&+\log\binom{4^L-l_2q_2}{M-l_2}-m\log\binom{4^{L_1}}{\frac{M-l_2}{m}}\\
&+m\left(\log\binom{4^{L_1}}{\frac{M-l_2}{m}}-\log\binom{4^{L_1}-l_1q_1}{\frac{M-l_2}{m}-l_1}\right)\\
\overset{(a)}{<}&l_2(2L+1)+ml_1(2L_1+1)\\
&+\sum_{i=1}^{M-l_2}\log(4^L-l_2q_2+1-i)-\log(M-l_2)!\\
&-m\sum_{i=1}^{\frac{M-l_2}{m}}\log(4_1^L-i+1)+m\log(\frac{M-l_2}{m})!\\
<&l_2(2L+1)+ml_1(2L_1+1)\\
&+m\log(\frac{M-l_2}{m})!-\log(M-l_2)!\\
&+m\sum_{i=1}^{\frac{M-l_2}{m}}\log\frac{4^L-l_2q_2-(i-1)m}{4^{L_1}-i+1}\\
<&l_2(2L+1)+ml_1(2L_1+1)+(M-l_2)2t\\
&+m\sum_{i=1}^{\frac{M-l_2}{m}}\log\left(1+\frac{(i-1)(4^t-m)-l_2q_2}{4^L-(i-1)\cdot 4^t}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+m\left((\frac{M-l_2}{m}+\frac{1}{2})\log\frac{M-l_2}{m}-\frac{M-l_2}{m}\log e+\log e\right)\\
&-(M-l_2+\frac{1}{2})\log(M-l_2)+(M-l_2)\log e-\log\sqrt{2\pi}\\
<&l_2(2L+1)+ml_1(2L_1+1)+(M-l_2)2t\\
&+m\sum_{i=1}^{\frac{M-l_2}{m}}\frac{(i-1)(4^t-m)-l_2q_2}{4^L-(i-1)\cdot 4^t}\\
&+\frac{m-1}{2}\log(M-l_2)-(M-l_2+\frac{m}{2})\log m+m\log e\\
<&l_2(2L+1)+ml_1(2L_1+1)+(M-l_2)(2t-\log m)\\
&+m\sum_{i=1}^{\frac{M-l_2}{m}}\frac{i-1}{\frac{M-l_2}{m}(q_1+1)-l_1-i}+\frac{m}{2}\log\frac{e^2(M-l_2)}{m}\\
<&l_2(2L+1)+ml_1(2L_1+1)+(M-l_2)(2t-\log m)\\
&+m\sum_{i=1}^{\frac{M-l_2}{m}}\frac{2i-1}{\frac{M-l_2}{m}q_1}+(M-l_2)\frac{1}{2\frac{M-l_2}{m}}\log\frac{e^2(M-l_2)}{m}\\
<&l_2(2L+1)+ml_1(2L_1+1)+(M-l_2)(2t-\log m)\\
&+m\left(\frac{M-l_2}{m}\right)^2\frac{1}{\frac{M-l_2}{m}q_1}+(M-l_2)\frac{1}{8}\log(4e^2)\\
<&l_2(2L+1)+ml_1(2L_1+1)+M(2t-\log m+1).
\end{aligned}
$$

Here (a) is true by applying Theorem 1 to this lemma. ■

Although Construction 2 has low redundancy, it has two other drawbacks. First, the mapping of the information to the codewords does not preserve the distance. A single edit in the codeword might result in a huge deviation from the original information, and a single bit change in the original file will also require rewriting a large number of strands. In the next section, we therefore focus on linear block codes, where the information can be stored by a systematic code, which are easier to rewrite and to decode.

IV. Locally Recoverable Linear Block Codes

In this section, we study the locality of linear block codes. Suppose the first t symbols in each DNA strand form an unique primer, and the remaining $L_1 = L - t$ symbols contain the original information. Then, the order of the subsets is determined by the primers; the set of codewords can be regarded as a code of length M in $GF(4^{L_1})$.

Definition 4. *A linear block code $\mathcal{C}$ defined on $GF(q)$ is called an $(m, n, k, d_1, d_2)_q$-code if $\mathcal{C}$ maps the information vector of mk symbols into codewords of mn symbols with the form $\boldsymbol{c} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_m)$, where $\boldsymbol{c}_m \in GF(q)^n$, with an overall minimum distance d_2 and a minimum distance d_1 of each $\boldsymbol{c}_i$.*

Lemma 4 is a known result from e.g., [18], which provides an upper bound of the global minimum distance of a code for a fixed local minimum distance. While the construction in [18] is indeed locally recoverable, it is not systematic. We provide in Construction 3 a general construction of locally recoverable codes that is systematic, both locally and globally, and which is more appropriate to be applied in DNA storage.

Lemma 4. *For an $(m, n, k, d_1, d_2)_q$-code, let $r = n - k$. For $\delta \in \mathbb{N}$, $\delta < r$, if $d_2 \leq n + 1$, $d_1 = r - \delta + 1$, then $d_2 \leq r + (m - 1)\delta + 1$.*

Definition 5. *A matrix $X \in GF(q)^{u\times v}$ is called a **good** matrix if every $k \times k$ submatrix of X, $1 \leq k \leq \min\{u, v\}$, is nonsingular.*

Lemma 5. *Suppose $a_1, \cdots, a_u, b_1, \cdots, b_v$ are pairwise distinct elements in $GF(q)$, then the following matrix is a **good** matrix,*

$$
\begin{bmatrix}
\frac{1}{a_1-b_1} & \frac{1}{a_1-b_2} & \cdots & \frac{1}{a_1-b_v}\\
\frac{1}{a_2-b_1} & \frac{1}{a_2-b_2} & \cdots & \frac{1}{a_2-b_v}\\
\vdots & \vdots & \ddots & \vdots\\
\frac{1}{a_u-b_1} & \frac{1}{a_u-b_2} & \cdots & \frac{1}{a_u-b_v}
\end{bmatrix}.
$$

Proof. For $1 \leq k \leq \min\{u, v\}$, denote the submatrix generated by the intersection of the $i_1, \cdots, i_k$-th rows and the $j_1, \cdots, j_k$-th columns by A. Then,

$$
A = \begin{bmatrix}
\frac{1}{a_1-b_1} & \frac{1}{a_1-b_2} & \cdots & \frac{1}{a_1-b_k}\\
\frac{1}{a_2-b_1} & \frac{1}{a_2-b_2} & \cdots & \frac{1}{a_2-b_k}\\
\vdots & \vdots & \ddots & \vdots\\
\frac{1}{a_k-b_1} & \frac{1}{a_k-b_2} & \cdots & \frac{1}{a_k-b_k}
\end{bmatrix}.
$$

We know that

$$
det(A) = \frac{\prod\limits_{1\leq i<j\leq n}(a_i - a_j)\prod\limits_{1\leq i<j\leq n}(b_j - b_i)}{\prod\limits_{i,j=1}^{k}(a_i - b_j)} \neq 0 \in GF(q).
$$

Therefore, A is nonsingular. ■

Lemma 6. *Let $m, n, p, r \in \mathbb{N}$, $m > n > r$, $A \in GF(q)^{m\times n}$, and $B \in GF(q)^{m\times p}$. Define matrix C, D, E as follows. If A is a good matrix, then any n rows from each matrix are all linearly independent.*

$$
C = \begin{bmatrix} A \\ -I_n \end{bmatrix}, D = \begin{bmatrix} A \\ -I_r\ \mathbf{0}_{n-r} \end{bmatrix}, E = \begin{bmatrix} A\ B \\ -I_{n+p} \end{bmatrix}.
$$

Proof. Suppose there exist n rows from C that are linearly dependent. Suppose a of these linearly dependent rows $r_1, \cdots, r_a$ are from A, and the other $n - a$ rows $r_{a+1}, \cdots, r_n$ are from $-I_n$, where $0 < a \leq n$. Suppose the entries of -1 in $r_{a+1}, \cdots, r_n$ are in the $j_1, \cdots, j_{n-a}$-th columns of C. Suppose $[n] \setminus \{j_1, \cdots, j_{n-a}\} = \{c_1, \cdots, c_a\}$. Then the $a \times a$ submatrix generated by the intersection of the rows $r_1, \cdots, r_a$ and the $c_1, \cdots, c_a$-th columns of A is singular. A contradiction!

Suppose there exist n rows from D that are linearly dependent. Suppose a of these linearly dependent rows $r_1, \cdots, r_a$ are from A, and the other $n - a$ rows $r_{a+1}, \cdots, r_n$ are from $[-I_r\ \mathbf{0}_{n-r}]$, where $n - r \leq a \leq n$. Suppose the entries of -1 in $r_{a+1}, \cdots, r_n$ are in the $j_1, \cdots, j_{n-a}$-th columns of C, then $j_k \leq r$ for all $1 \leq k \leq r$. Suppose $[n]\setminus\{j_1, \cdots, j_{n-a}\} = \{c_1, \cdots, c_a\}$. Then the $a\times a$ submatrix generated by the intersection of the rows $r_1, \cdots, r_a$ and the $c_1, \cdots, c_a$-th columns of A is singular. A contradiction!

Suppose there exist n rows from E that are linearly dependent. Suppose a of these linearly dependent rows $r_1, \cdots, r_a$ are from the first $m+n$ rows of E, $n-a$ rows $r_{a+1}, \cdots, r_{n-a}$ are from $[\mathbf{0}_n \ \ -\boldsymbol{I}_p]$, where $0 < a \leq n$. From the previous discussion of matrix C, we know that $a \geq n+1$, which means that $n > a \geq n+1$. A contradiction! ∎

Construction 3. *Let $n, k, m, r, \delta \in \mathbb{N}$, $r = n-k$, $k > r + (m-1)\delta$, $0 < \delta < r$, and suppose $A_{i,j} \in GF(q)^{k \times n}$ for all $1 \leq i, j \leq m$. Define $G \in GF(q)^{mk \times mn}$ as follows,*

$$G = \left[\begin{array}{c|c|c|c|c|c|c} I_k & A_{1,1} & \mathbf{0}_k & A_{1,2} & \cdots & \mathbf{0}_k & A_{1,m} \\ \hline \mathbf{0}_k & A_{2,1} & I_k & A_{2,2} & \cdots & \mathbf{0}_k & A_{2,m} \\ \hline \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hline \mathbf{0}_k & A_{m,1} & \mathbf{0}_k & A_{m,2} & \cdots & I_k & A_{m,m} \end{array}\right].$$

Suppose $A_{i,j} = B_{i,j} X_{i,j} U_j$, $1 \leq i, j \leq m$, $i \neq j$, for some $B_{i,j} \in GF(q)^{k \times \delta}$, $X_{i,j} \in GF(q)^{\delta \times \delta}$, $U_j \in GF(q)^{\delta \times r}$, such that the following conditions are satisfied:

1) *$rank(B_{i,j}) = rank(U_j) = rank(X_{i,j}) = \delta$, where $rank(\cdot)$ refers to the rank of the matrix in $GF(q)$;*
2) *$[A_{i,i}, B_{i,1}, \cdots, B_{i,m}]$, $1 \leq i \leq m$, are good matrices;*
3) *$\left[A_{i,i}^T, U_i^T\right]$, $1 \leq i \leq m$, are good matrices.*

Then, G is the generator matrix of an $(m, n, k, r-\delta+1, d)_q$-code, where $d \geq \min\{r + (m-1)\delta + 1, 2(r-\delta+1)\}$.

Moreover, denote the matrix consisting of the first δ columns of $A_{j,j}$ by $B_{j,j}$, and that consisting of the first δ columns of U_j by C_j. If the following matrix B is good and $X_{i,j} = C_j^{-1}$, then $d \geq \min\{r + (m-1)\delta + 1, \max\{2(r-\delta+1), m\delta+1\}\}$,

$$B = \begin{bmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,m} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m,1} & B_{m,2} & \cdots & B_{m,m} \end{bmatrix}.$$

Proof. First we prove that $d_1 = r - \delta + 1$, i.e., every subblock has minimum Hamming distance $d_1 = r - \delta + 1$. Suppose the parity check matrix for the j-th block is H_j, $1 \leq j \leq m$. Then the following equation follows,

$$[I_k \ A_{j,j}] \, H_j = 0, [0 \ A_{i,j}] \, H_j = 0.$$

Therefore,

$$H_j = \begin{bmatrix} A_{j,j} \\ -I_r \end{bmatrix} V_j,$$

for some $V_j \in GF(q)^{r \times t}$, $rank(V_j) = t$, $t \leq r$. From $[0, A_{i,j}] \, H_j = 0$, we know that $B_{i,j} X_{i,j} U_j V_j = 0$. Given that $rank(B_{i,j}) = rank(X_{i,j}) = \delta$, and $B_{i,j} X_{i,j} \in GF(q)^{k \times \delta}$, we have $U_j V_j = 0$ and $\mathcal{R}(U_j) = \mathcal{N}(V_j^T)$.

Suppose $\boldsymbol{u}_j H_j = 0$ for some $\boldsymbol{u}_j \in GF(q)^n$, $\boldsymbol{u}_j \neq 0$, we know that

$$0 = \boldsymbol{u}_j \begin{bmatrix} A_{j,j} \\ -I_r \end{bmatrix} V_j \Longrightarrow \boldsymbol{u}_j \begin{bmatrix} A_{j,j} \\ -I_r \end{bmatrix} = \boldsymbol{s}_j U_j,$$

for some $\boldsymbol{s}_j \in GF(q)^\delta$.

Therefore,

$$0 = [\boldsymbol{u}_j, -\boldsymbol{s}_j] \begin{bmatrix} A_{j,j} \\ -I_r \\ U_j \end{bmatrix}.$$

Then from Lemma 6, $w_H(\boldsymbol{u}_j) + w_H(\boldsymbol{s}_j) = w_H([\boldsymbol{u}_j, -\boldsymbol{s}_j]) \geq r + 1$. Given that $w_H(\boldsymbol{s}_j) \leq \delta$, we have $w_H(\boldsymbol{u}_j) \geq r - \delta + 1$, which means that $d_1 \geq r - \delta + 1$.

Moreover, since the matrix $E_j^T = \left[A_{j,j}^T, U_j^T\right]$ is good, any r different rows of E_j are linearly independent. Therefore consider the matrix F_j containing the last $r+1$ rows of E_j, we know that there exists a vector $\boldsymbol{x}_j = [\boldsymbol{v}_j, \boldsymbol{h}_j]$, where $\boldsymbol{v}_j \in GF(q)^{r-\delta+1}$, $\boldsymbol{h}_j \in GF(q)^\delta$, such that $\boldsymbol{x}_j F_j = 0$, and $\boldsymbol{x}_j$ has no zero entries. Let $\boldsymbol{u}_j = \left[0^{k-r+\delta-1}, \boldsymbol{v}_j, 0^r\right]$, then,

$$[\boldsymbol{u}_j, \boldsymbol{h}_j] \begin{bmatrix} A_{j,j} \\ -I_r \\ U_j \end{bmatrix} = \boldsymbol{x}_j F_j = 0,$$

which means that

$$\boldsymbol{u}_j \begin{bmatrix} A_{j,j} \\ -I_r \end{bmatrix} V_j = -\boldsymbol{h}_j U_j V_j = 0,$$

which means that $\boldsymbol{u}_j$ is a nonzero codeword of the j-th subblock, and $w_H(\boldsymbol{u}_j) = w_H(\boldsymbol{v}_j) = r - \delta + 1$. Therefore $d_1 = r - \delta + 1$.

Secondly, we prove that $d \geq \min\{r + (m-1)\delta + 1, 2(r - \delta + 1)\}$. Suppose $\boldsymbol{u} = (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m)$ is a nonzero codeword, for some $\boldsymbol{u} \in GF(q)^{mn}$, $\boldsymbol{u}_j \in GF(q)^n$, and for all $1 \leq j \leq m$. Suppose the parity check matrix of this code is H. Define H as follows:

$$H = \left[\begin{array}{c|c|c|c} A_{1,1} & A_{1,2} & \cdots & A_{1,m} \\ \hline -I_r & \mathbf{0}_r & \cdots & \mathbf{0}_r \\ \hline A_{2,1} & A_{2,2} & \cdots & A_{2,m} \\ \hline \mathbf{0}_r & -I_r & \cdots & \mathbf{0}_r \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline A_{m,1} & A_{m,2} & \cdots & A_{m,m} \\ \hline \mathbf{0}_r & \mathbf{0}_r & \cdots & -I_r \end{array}\right].$$

Suppose $\exists j$, $1 \leq j \leq m$, such that $\boldsymbol{u}_j \neq 0$, and $\boldsymbol{u}_i = 0$, for all $1 \leq i \leq m$, $i \neq j$. Then $\boldsymbol{u}_j$ satisfies that:

$$\boldsymbol{u}_j \begin{bmatrix} A_{j,j} \\ -I_r \end{bmatrix} = 0, \boldsymbol{u}_j \begin{bmatrix} A_{i,j} \\ \mathbf{0}_r \end{bmatrix} = 0 \Longleftrightarrow \boldsymbol{u}_j \begin{bmatrix} B_{i,j} \\ \mathbf{0}_\delta \end{bmatrix} = 0.$$

Therefore,

$$\boldsymbol{u}_j \begin{bmatrix} A_{j,j} & B_{1,j} & \cdots & B_{j-1,j} & B_{j+1,j} & \cdots & B_{m,j} \\ -I_r & & & \mathbf{0}_{(m-1)\delta} & & & \end{bmatrix} = 0.$$

From Lemma 6, we know that $w_H(\boldsymbol{u}) = w_H(\boldsymbol{u}_j) \geq r + (m-1)\delta + 1$.

Suppose $\exists i, j$, $1 \leq i < j \leq m$, such that $\boldsymbol{u}_i \neq 0$, and $\boldsymbol{u}_j \neq 0$. Then $w_H(\boldsymbol{u}) \geq w_H(\boldsymbol{u}_i) + w_H(\boldsymbol{u}_j) \geq 2d_1 = 2(r - \delta + 1)$.

Therefore, $d = \min w_H(\boldsymbol{u}) \geq \min\{r + (m-1)\delta + 1, 2(r - \delta + 1)\}$.

Lastly, we prove that $d \geq m\delta + 1$ given that B is good and $X_{i,j} = C_j^{-1}$ in $GF(q)$, where C_j is the matrix consisting of the first δ columns of U_j.

Rearrange the rows and columns of H to obtain $\tilde{H}$, namely, $\exists$ permutation matrices P, Q such that $PHQ = \tilde{H}$, where

$$\tilde{H} = \begin{bmatrix} B & A \\ -I_{m\delta} & \mathbf{0}_{m(r-\delta)} \\ \mathbf{0}_{m\delta} & -I_{m(r-\delta)} \end{bmatrix}.$$

Then for every nonzero codeword $\boldsymbol{u}$, let $\tilde{\boldsymbol{u}} = \boldsymbol{u}P^{-1}$, we have $\tilde{\boldsymbol{u}}\tilde{H} = 0$ and $\tilde{\boldsymbol{u}} \neq 0$. From Lemma 6, $w_H(\boldsymbol{u}) = w_H(\tilde{\boldsymbol{u}}) \geq m\delta + 1$. ■

Theorem 3. *Suppose $a_1, \cdots, a_{mk+\delta}, b_1, \cdots, b_{mr}$ are pairwise distinct elements in $GF(q)$. Define $A_{j,j}$, $B_{i,j}$, U_j, $X_{i,j}$ as follows, for all $1 \leq i, j \leq m$, $i \neq j$, in Construction 3, then G is the generator matrix of an $(m, n, k, r - \delta + 1, d)_q$-code, where $d \geq \min\{r + (m-1)\delta + 1, \max\{2(r - \delta + 1), m\delta + 1\}\}$:*

$$\begin{aligned} A_{j,j} &= Y((j-1)k, jk, (j-1)r, jr), \\ B_{i,j} &= Y((i-1)k, ik, (j-1)r, (j-1)r + \delta), \\ U_j &= Y(mk, mk + \delta, (j-1)r, jr), \\ X_{i,j} &= Y(mk, mk + \delta, (j-1)r, (j-1)r + \delta)^{-1}, \end{aligned}$$

where the matrices $Y(i_1, i_2, j_1, j_2)$ for $0 \leq i_1 < i_2 \leq mk + \delta$, $0 \leq j_1 < j_2 \leq mr$ are defined as below,

$$Y(i_1, i_2, j_1, j_2) =$$
$$= \begin{bmatrix} \frac{1}{a_{i_1+1} - b_{j_1+1}} & \frac{1}{a_{i_1+1} - b_{j_1+2}} & \cdots & \frac{1}{a_{i_1+1} - b_{j_2}} \\ \frac{1}{a_{i_1+2} - b_{j_1+1}} & \frac{1}{a_{i_1+2} - b_{j_1+2}} & \cdots & \frac{1}{a_{i_1+2} - b_{j_2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{a_{i_2} - b_{j_1+1}} & \frac{1}{a_{i_2} - b_{j_1+2}} & \cdots & \frac{1}{a_{i_2} - b_{j_2}} \end{bmatrix}.$$

Proof. It follows immediately from Construction 3 and Lemma 5. ■

In Theorem 3, we proposed an explicit construction of a systematic $(m, n, k, r - \delta + 1, d)_q$-code, where $d \geq \min\{r + (m-1)\delta + 1, \max\{2(r - \delta + 1), m\delta + 1\}\}$. This code has two properties that make it suitable for rewritable random-access DNA storage. First, the code is systematic, which means that a single edit of the message results in rewriting at most mr DNA strands. Second, each information vector can be derived directly from the corresponding subblock when there are at most $r - \delta$ local errors within that subblock. When $\delta \leq \frac{r+1}{m+1}$, the global minimum distance is $r + (m-1)\delta + 1$, which reaches the upper bound provided in Lemma 4. The following Theorem 4 proves the existence of a code that reaches the upper bound for any $\delta < r$, provided that the alphabet size is large enough. Future discussion includes finding explicit constructions that reach the upper bound for arbitrary $\delta < r$.

Theorem 4. *For any $\delta, m, n, k \in \mathbb{N}$, let $r = n - k$, and suppose $\delta < r$. If $q > mn + \delta + (r + (m-1)\delta - 1)\binom{mn-1}{r+(m-1)\delta} - (mr - 1)\binom{mn-2}{r+(m-1)\delta-1}$, then there exist pairwise distinct elements $\alpha_1, \cdots, \alpha_{mk+\delta}, b_1, \cdots, b_{mr}$ from $GF(q)$ such that Theorem 3 presents a generator matrix of an $(m, n, k, r - \delta + 1, r + (m-1)\delta + 1)_q$-code.*

Proof. The condition $d_2 \geq r + (m-1)\delta + 1$ is true if every $(r + (m-1)\delta - a) \times (mr - a)$ submatrix of the submatrix $A(s)$ has linearly independent rows, for all $1 \leq s \leq mr$, where $A(s)$ refers to the submatrix consisting of the first s columns of the following matrix A,

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,m} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,m} \end{bmatrix}.$$

First, choose arbitrary pairwise distinct elements $a_1, a_2, \cdots, a_{mk+\delta}$, and $b_1, \cdots, b_{(m-1)(r-\delta)}$, from $GF(q)$. Suppose $b_1, \cdots, b_s$, $(m-1)(r-\delta) \leq s < mr$ are determined, and every $(r + (m-1)\delta - a) \times (mr - a)$ submatrix of the submatrix $A(s+1)$ has linearly independent rows. Then all the elements in the $s+1$-th column are either $\frac{1}{a_i - b_{s+1}}$, $1 \leq i \leq mk$, or a linear combination of $\frac{1}{a_i - b_{s+1}}$, $mk + 1 \leq i \leq mk + \delta$, and every $(r + (m-1)\delta - a) \times (mr - a)$ submatrix has the following form:

$$C = \begin{bmatrix} \boldsymbol{b}_1 & f_1(b_{s+1}) \\ \boldsymbol{b}_2 & f_2(b_{s+1}) \\ \vdots & \vdots \\ \boldsymbol{b}_{r+(m-1)\delta-a} & f_{r+(m-1)\delta-a}(b_{s+1}) \end{bmatrix},$$

where $\boldsymbol{b}_i \in GF(q)^{mr-a-1}$, and any $r + (m-1)\delta - a - 1$ rows from $\{\boldsymbol{b}_i\}_{i=1}^{r+(m-1)\delta-a}$ are linearly independent; and $f_i(b_{s+1})$, $1 \leq i \leq r + (m-1)\delta - a$, is either $\frac{1}{a_i - b_{s+1}}$, $1 \leq i \leq mk$, or a linear combination of $\frac{1}{a_i - b_{s+1}}$, $mk + 1 \leq i \leq mk + \delta$. Then if C has linearly dependent rows, i.e., $\exists \boldsymbol{b} \in GF(q)^{r+(m-1)\delta-a}$, such that $\boldsymbol{b}C = 0$. We know that the nullspace of $\left[\boldsymbol{b}_1^T, \boldsymbol{b}_2^T, \cdots, \boldsymbol{b}_{r+(m-1)\delta-a}^T\right]$ has dimension at most 1, it could only be $\{\gamma\boldsymbol{b}, \gamma \in GF(q)\}$. Suppose $\boldsymbol{b} = (\beta_1, \cdots, \beta_{r+(m-1)\delta-a})$, then,

$$\sum_{i=1}^{r+(m-1)\delta-a} \beta_i f_i(b_{s+1}) = 0,$$

which means that b_{s+1} is a root of a polynomial of degree less than $r + m\delta - a$, and thus there are at most $r + m\delta - a - 1$ elements such that C has linearly dependent rows. Since there are $\binom{mk}{r+(m-1)\delta-a}\binom{s}{mr-a-1}$ choices of C, there exists a b_{s+1} such that any $(r + (m-1)\delta - a) \times (mr - a)$ submatrix of the following matrix A has linearly independent rows as long as:

$$q > mk + \delta + s + \sum_{a=0}^{r+(m-1)\delta} (r + (m-1)\delta - a - 1)$$
$$\binom{mk}{r + (m-1)\delta - a}\binom{s}{mr - a - 1}.$$

Therefore, given the following inequality, there exist pairwise distinct elements $\alpha_1, \cdots, \alpha_{mk+\delta}, b_1, \cdots, b_{mr}$ from $GF(q)$ such that Theorem 3 presents a generator matrix of an $(m, n, k, r-\delta+1, r+(m-1)\delta+1)$-code:

$$\begin{aligned}
q \geq & mk+\delta+mr+\sum_{a=0}^{r+(m-1)\delta}(r+(m-1)\delta-a-1) \\
& \binom{mk}{r+(m-1)\delta-a}\binom{mr-1}{mr-a-1} \\
= & mn+\delta+(r+(m-1)\delta-1)\binom{mn-1}{r+(m-1)\delta} \\
& -(mr-1)\binom{mn-2}{r+(m-1)\delta-1}.
\end{aligned}$$

■

Example 2. *Suppose $M=4$, $m=2$, $r=2$, $\delta=1$, $k=3$, $n=k+r=5$. Then $mk+\delta=7$ and $mr=4$. Suppose $q=16>7+4$. Let $a_i=i$, $1 \leq i \leq 7$, $b_j=7+j$, $1 \leq j \leq 4$. Therefore the following $G \in GF(q)^{6\times 10}$ is the generating matrix of a $(2,5,3,2,4)_{16}$-code.*

$$G=\left[\begin{array}{ccc|cc|ccc|cc}
1 & 0 & 0 & 2 & 15 & 0 & 0 & 0 & 1 & 11 \\
0 & 1 & 0 & 12 & 5 & 0 & 0 & 0 & 6 & 15 \\
0 & 0 & 1 & 5 & 12 & 0 & 0 & 0 & 11 & 9 \\
\hline
0 & 0 & 0 & 6 & 12 & 1 & 0 & 0 & 3 & 8 \\
0 & 0 & 0 & 3 & 6 & 0 & 1 & 0 & 8 & 3 \\
0 & 0 & 0 & 7 & 14 & 0 & 0 & 1 & 10 & 4
\end{array}\right].$$

V. Conclusion

In this paper, we studied codes for DNA storage systems from two aspects. We first presented a class of rate-efficient codes that have redundancy up to a constant factor times the optimal redundancy. Then, we extended our construction to be locally recoverable. While these constructions are rate-efficient, they are not distance preserving and they have high complexity in the decoding and rewriting processes, especially compared to linear block codes. We then discussed locally recoverable linear block codes. We proved the existence of such codes where the global minimum distance reaches the upper bound, given a prescribed local minimum distance and a fixed redundancy, for a large enough alphabet size. Future work will focus on designing explicit locally recoverable linear block codes or LDPC codes that can be used in DNA storage systems.

Acknowledgment

The authors would like to thank Raghu Meka from UCLA for helpful discussions, and Dana Watson for the suggestions on writing. This work has received funding from DYF, NSF under the grant CCF-BSF 1718389, and from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n. PCOFUND-GA-2013-609102.

References

[1] V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church, and W. L. Hughes, "Nucleic acid memory," *Nature materials*, vol. 15, no. 4, p. 366, 2016.

[2] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, no. 6012, 2012.

[3] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen *et al.*, "Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, Mar. 2018.

[4] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3125–3146, Jun. 2016.

[5] M. Blawat, K. Gaedke, I. Huetter, X.-M. Chen, B. Turczyk, S. Inverso, B. W. Pruitt, and G. M. Church, "Forward error correction for DNA data storage," *Procedia Computer Science*, vol. 80, pp. 1011–1022, 2016.

[6] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *arXiv preprint arXiv:1803.03322*, 2018.

[7] R. Gabrys, H. M. Kiah, and O. Milenkovic, "Asymmetric Lee distance codes for DNA-based storage," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 4982–4995, Aug. 2017.

[8] R. Gabrys, E. Yaakobi, and O. Milenkovic, "Codes in the Damerau distance for deletion and adjacent transposition correction," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2550–2570, April 2018.

[9] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage," *arXiv preprint arXiv:1801.04882*, 2018.

[10] S. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Scientific reports*, vol. 5, p. 14138, 2015.

[11] M. Levy and E. Yaakobi, "Mutually uncorrelated codes for DNA storage," June 2017, pp. 3115–3119.

[12] S. Jain, F. F. Hassanzadeh, M. Schwartz, and J. Bruck, "Duplication-correcting codes for data storage in the DNA of living organisms," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 4996–5010, Aug. 2017.

[13] Y. M. Chee, J. Chrisnata, H. M. Kiah, and T. T. Nguyen, "Efficient encoding/decoding of irreducible words for codes correcting Tandem duplications," *arXiv preprint arXiv:1801.02310*, 2018.

[14] J. Shendure, S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, J. A. Schloss, and R. H. Waterston, "DNA sequencing at 40: past, present and future," *Nature*, vol. 550, no. 7676, p. 345, Oct. 2017.

[15] T. L. Schmidt, B. J. Beliveau, Y. O. Uca, M. Theilmann, F. Da Cruz, C.-T. Wu, and W. M. Shih, "Scalable amplification of strand subsets from chip-synthesized oligonucleotide libraries," *Nature communications*, vol. 6, p. 8634, 2015.

[16] M. Schirmer, R. DAmore, U. Z. Ijaz, N. Hall, and C. Quince, "Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data," *BMC bioinformatics*, vol. 17, no. 1, p. 125, 2016.

[17] S. Yang, C. Schoeny, and L. Dolecek, "Order-optimal permutation codes in the generalized Cayley metric," in *IEEE Information Theory Workshop*, Kaohsiung, Taiwan, Nov 2017, pp. 234–238.

[18] J. Han and L. A. Lastras-Montano, "Reliable memories with subline accesses," in *Proc. IEEE Int. Symp. Inf. Theory*, June 2007, pp. 2531–2535.