Sequence-Subset Distance and Coding for Error Control in DNA-based Data Storage

Wentu Song and Kui Cai

Abstract

The process of DNA-based data storage (DNA storage for short) can be mathematically modelled as a communication channel, termed DNA storage channel, whose inputs and outputs are sets of unordered sequences. To design error correcting codes for DNA storage channel, a new metric, termed the sequence-subset distance, is introduced, which generalizes the Hamming distance to a distance function defined between any two sets of unordered vectors and helps to establish a uniform framework to design error correcting codes for DNA storage channel. We further introduce a family of error correcting codes, referred to as sequence-subset codes, for DNA storage and show that the error-correcting ability of such codes is completely determined by their minimum distance. We derive some upper bounds on the size of the sequence-subset codes including a Singleton-like bound and a Plotkin-like bound. We also propose some constructions, which imply lower bounds on the size of such codes.

I. INTRODUCTION

The idea of storing data in synthetic DNA sequences has been around since 1988 [1] and DNA-based data storage has been progressing rapidly in recent years with the development of DNA synthesis and sequencing technology. Compared to traditional magnetic and optical media, DNA storage has competing advantages including extreme high density, long durability [7], and low energy consumption [2].

A DNA sequence is mathematically represented by a quaternary sequence, each symbol represent one of the four types of base nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). Basically, in a DNA-based storage system, the original binary data is first encoded to a set of quaternary sequences. Then the corresponding DNA nucleotide sequences (oligos) are synthesized and stored. To retrieve the original data, the stored oligos are sequenced to generate a set of quaternary sequences, which then are decoded to the original binary data. The process of DNA synthesizing, storing and sequencing can be mathematically modelled as a communication channel, called the DNA storage channel, which can be depicted by Fig. 1.

![Fig 1. System model of the DNA storage: The DNA storage channel is the mathematical model of the process of DNA synthesizing, storing and sequencing. A reliable system should guarantee that with sufficiently high probability the decoded file $\mathcal{F}$ equals to the original file $\tilde{\mathcal{F}}$.](image)

Unlike the conventional magnetic or optical recording systems, the DNA sequences are stored in “pools”, where structured addressing is not allowed. Therefore, the inputs and outputs of the DNA storage channel can only be viewed as sets of unordered DNA sequences.

The output of the DNA storage channel may be distorted by the following five types of errors:

- **Sequence deletion**: One or more of the input sequences are lost. As a result, the number of output sequences is smaller than the number of input sequences.
- **Sequence insertion**: One or more sequences that do not belong to the set of input sequences are added into the output sequences. As a result, the number of output sequences is larger than the number of input sequences.
- **Symbol deletion**: One or more symbols in a sequence are removed. As a result, the length of the erroneous sequence is decreased.
- **Symbol insertion**: One or more symbols are added into a sequence. As a result, the length of the erroneous sequence is increased.
- **Symbol substitution**: One or more symbols in a sequence are replaced by other symbols. In this case, the length of the erroneous sequence remains unchanged.

Note that sequence deletion and sequence insertion can take place simultaneously. If the number of sequence deletions equals the number of sequence insertions, then the total number of input sequences remain unchanged. In this case, the combining effect of sequence deletion and sequence insertion is equivalent to symbol substitutions.

Wentu Song and Kui Cai are with Singapore University of Technology and Design, Singapore (e-mails: {wentusong, cai_kui}@sutd.edu.sg).
To combat different types of errors in DNA synthesizing and sequencing, various coding techniques are used by DNA storage. Most demonstration researches employ constrained coding combined with classical error correcting codes (e.g. Reed-Solomon codes) \cite{2,10}. In addition, to combat the lack of ordering of the transmitted sequences, a unique address (index) is added to each sequence.

Codes that can correct \(s\) (or fewer) losses of sequences and \(e\) (or fewer) substitutions in each of \(t\) (or fewer) sequences were studied in \cite{11} by considering the so-called error ball. Codes dealing with insertion/deletion errors were also studied in \cite{11}. Codes that can correct a total of \(K\) substitution errors were studied in \cite{12} using the sphere packing arguments, which is essentially the same as the error ball arguments.

### A. Our Contribution

In this paper, we consider error control for DNA storage channel by introducing a new metric, termed the sequence-subset distance, over the power set of the set of all vectors over a finite alphabet with fixed length, which is the space of the inputs/outputs of the DNA storage channel. This metric is a generalization of the classical Hamming distance and can help to establish a uniform framework to design codes for DNA storage channel that can correct errors of sequence deletion, sequence insertion and symbol substitution.

We study error correcting codes with respect to the sequence-subset distance, which we refer to as sequence-subset codes, for DNA-based data storage. We show that similar to codes with respect to the classical Hamming distance, a sequence-subset code \(C\) can correct any number of \(n_D\) sequence deletions, \(n_I\) sequence insertions, and \(n_S\) symbol substitutions, provided
\[
\sum_{i=0}^{n_S} L \cdot \max\{n_i, n_D\} \leq \frac{d(s(C))}{M_0},
\]
where \(L\) is the length of the sequences and \(d(s(C))\) is the minimum distance of \(C\).

We derived some upper bounds on the size of the sequence-subset codes including a Singleton-like bound and a Plotkin-like bound.

We give a construction of optimal codes (with respect to size) for the special case that \(L \mid d\) and \(M_0 = \frac{d}{L}\). We also give some general constructions, which imply lower bounds of the size of such codes.

### B. Other Related Work

In \cite{13}, the input and output of the DNA storage channel are both viewed as a multi-set, rather than set, of DNA molecules, where the numbers of the input sequences and output sequences may be different. The fundamental limits of the DNA storage model was investigated under the assumption that each sampled molecule is read in an error-free manner.

Another different channel model for DNA storage was studied in \cite{14}, where the process of DNA storage is modelled by two successive channels, i.e., the synthesis channel and the sequencing channel, and the output of the sequencing channel is a set of DNA fragments which can be represented by a profile vector. And three types of errors, namely, substitution errors due to synthesis, coverage errors, and \(\ell\)-gram substitution errors due to sequencing, are considered.

There are other communication channels similar to DNA storage channel. The permutation channel considered in \cite{15} has input and output as vectors over a finite alphabet and the transmitted vector is corrupted by a permutation on its coordination. Permutation channel with impairments was considered in \cite{16}, where the input and output are multi-sets, rather than vectors, of symbols from a finite alphabet. Such models are not appropriate for DNA storage because in these models, sequences are treated in the symbol level and the structure information of sequences is neglected.

### C. Organization

The rest of the paper is organized as follows. In Section II, we introduce the sequence-subset distance and provide the basic properties of codes with sequence-subset distance. We analyze the upper bound on the size of sequence-subset codes in Section III and give some constructions of such codes in Section IV. The paper is concluded in Section V.

### D. Notations

The following notations will be used in this paper:

1) For any positive integer \(n\), \([n] := \{1, 2, \ldots, n\}\).

2) For any set \(A\), \(|A|\) denotes the size (i.e., cardinality) of \(A\) and \(P(A)\) denotes the power set of \(A\) (i.e., the collection of all subsets of \(A\)).

3) For any two sets \(X\) and \(Y\), \(X \setminus Y\) is the set of all elements of \(X\) that do not belong to \(Y\).

4) For any \(n\)-tuple \(x \in A^n\) and any \(i \in [n]\), \(x(i)\) denotes the \(i\)th coordinate of \(x\), and hence \(x\) is denoted as \(x = (x(1), x(2), \ldots, x(n))\).
A. Sequence-Subset Distance

Let $A$ be a fixed finite alphabet. For DNA data storage, typically $A = \{A, T, C, G\}$, representing the four types of base nucleotides. In this work, for generality, we assume that $A$ is any fixed finite alphabet of size $q \geq 2$.

Let $L$ be a positive integer. For any $x_1, x_2 \in A^L$, the Hamming distance between $x_1$ and $x_2$, denoted by $d_H(x_1, x_2)$, is defined as the number of coordinates where $x_1$ and $x_2$ differ, that is,

$$d_H(x_1, x_2) := |\{i \in [L]; x_1(i) \neq x_2(i)\}|.$$

For any two subsets $X_1$ and $X_2$ of $A^L$ such that $|X_1| \leq |X_2|$ and any injection $\chi : X_1 \rightarrow X_2$, denote

$$d_\chi(X_1, X_2) := \sum_{x \in X_1} d_H(x, \chi(x)) + L(|X_2| - |X_1|).$$

Then a natural way to generalize Hamming distance to the space of all subsets of $A^L$ is as follows.

Definition 1: For any $X_1, X_2 \subseteq A^L$, without loss of generality, assuming $|X_1| \leq |X_2|$, the sequence-subset distance between $X_1$ and $X_2$ is defined as

$$d_S(X_1, X_2) = d_S(X_2, X_1) := \min_{\chi \in \mathcal{X}} d_\chi(X_1, X_2),$$

where $\mathcal{X}$ is the set of all injections $\chi : X_1 \rightarrow X_2$.

We first prove some important properties of the function $d_S(\cdot, \cdot)$. Then we will prove that it is really a distance function.

First, intuitively, the elements in $X_1 \cap X_2$ should have no effect on the sequence-subset distance between $X_1$ and $X_2$. This is shown to be true by the following lemma and corollary.

Lemma 1: For any $X_1, X_2 \subseteq A^L$ such that $|X_1| \leq |X_2|$, there exists a $\chi_0 \in \mathcal{X}$ such that $d_S(X_1, X_2) = d_{\chi_0}(X_1, X_2)$ and $\chi_0(x) = x$ for all $x \in X_1 \cap X_2$.

Proof: The proof is given in Appendix A.

Corollary 1: For any two subsets $X_1$ and $X_2$ of $A^L$,

$$d_S(X_1, X_2) = d_S(X_1 \setminus X_2, X_2 \setminus X_1).$$

Proof: This corollary is just a direct consequence of Definition 1 and Lemma 1.

Lemma 2: Suppose $X_1, X_2 \subseteq A^L$ such that $|X_1| \leq |X_2|$. Suppose $X_1', X_2' \subseteq X_2$ such that $|X_1| \leq |X_2'|$. Then

$$d_S(X_1, X_2') \leq d_S(X_1, X_2).$$

Proof: The proof is given in Appendix B.

Now we prove that $d_S(\cdot, \cdot)$ is really a distance function (metric) over $P(A^L)$.

Theorem 1: The function $d_S(\cdot, \cdot)$ is a distance function over the power set $P(A^L)$.

Proof: The proof is given in Appendix C.

B. Error Pattern of DNA Storage Channel

In this paper, we consider DNA storage channel with sequence deletion/insertion and symbol substitution. The input of the channel is a set of unordered sequences

$$X = \{x_1, x_2, \ldots, x_M\} \subseteq A^L$$

and the output is another set of unordered sequences

$$Y = \{y_1, y_2, \ldots, y_M\} \subseteq A^L,$$

where $L$ is the length of the sequences. Due to the channel noise, $Y \neq X$ is possible. Sequences in the subset $X \cap Y$ are correctly transmitted; Sequences in $X \setminus Y$ are either lost (sequence deletion) or changed to sequences in $Y \setminus X$ (symbol substitution); Sequences in $Y \setminus X$ are either excessive (sequence insertion) or obtained from some sequences in $X \setminus Y$ (symbol substitution). Let $n_1$, $n_0$, and $n_s$ denote the number of sequence insertions, sequence deletions, and symbol substitutions, respectively. Then we call the 3-tuple $(n_1, n_0, n_s)$ the error pattern of $Y$. And we have the following lemma.

Lemma 3: Suppose the channel input is $X$ and output is $Y$ of error pattern $(n_1, n_0, n_s)$. Then

$$d_S(X, Y) \leq n_s + L \cdot \max\{n_1, n_0\}.$$
Hence, similar to Case 1, we have

\[
X \setminus Y \subseteq X_\mathsf{D} \cup X_\mathsf{S} \quad \text{and} \quad Y \setminus X = Y_1 \cup Y_\mathsf{S},
\]

where \(X_\mathsf{D}\) is the set of lost input sequences, \(X_\mathsf{S}\) is the set of input sequences that are changed to \(Y_\mathsf{S}\) by symbol substitution, and \(Y_1\) is the set of sequences that are inserted to \(Y\). Clearly, we have

\[
n_1 = |Y_1| \quad \text{and} \quad n_\mathsf{D} = |X_\mathsf{D}|.
\]

Moreover, \(|X_\mathsf{S}| = |Y_\mathsf{S}|\) and there exists a bijection \(\chi : X_\mathsf{S} \rightarrow Y_\mathsf{S}\) such that for each \(x \in X_\mathsf{S}\), \(\chi(x)\) is the erroneous sequence of \(x\) by symbol substitution. Hence, we have

\[
n_\mathsf{S} = \sum_{x \in X_\mathsf{S}} d_H(x, \chi(x)).
\]

For further discussion, we need to consider the following two cases.

Case 1: \(n_1 \leq n_\mathsf{D}\). In this case, \(|Y_1| = n_1 \leq n_\mathsf{D} = |X_\mathsf{D}|\) and \(|Y \setminus X| \leq |X \setminus Y|\). So there exists an injection \(\chi' : Y_1 \rightarrow X_\mathsf{D}\) and we can let \(\bar{\chi} : Y \setminus X \rightarrow X \setminus Y\) be such that

\[
\bar{\chi}(y) = \begin{cases} 
\chi^{-1}(y) & \text{if } y \in Y_\mathsf{S}; \\
\chi'(y) & \text{if } y \in Y_1.
\end{cases}
\]

Since \(|X \setminus Y| - |Y \setminus X| = |X_\mathsf{D}| - |Y_1| = n_\mathsf{D} - n_1\), then by (1),

\[
d_{\bar{\chi}}(Y \setminus X, X \setminus Y) = \sum_{y \in Y \setminus X} d_H(y, \bar{\chi}(y)) + L \cdot (|X \setminus Y| - |Y \setminus X|)
\]

\[
= \sum_{y \in Y_\mathsf{S}} d_H(y, \chi(y)) + \sum_{y \in Y_1} d_H(y, \chi'(y)) + L \cdot (n_\mathsf{D} - n_1)
\]

\[
\leq n_\mathsf{S} + L \cdot n_1 + L \cdot (n_\mathsf{D} - n_1)
\]

\[
= n_\mathsf{S} + L \cdot n_\mathsf{D}
\]

\[
= n_\mathsf{S} + L \cdot \max\{n_1, n_\mathsf{D}\}
\]

where the inequality comes from the simple fact that \(d_H(z, z') \leq L\) for any \(z, z' \in \mathbb{A}^L\). Hence, by Corollary (1) and Definition (1) we have

\[
d_S(X, Y) = d_S(X \setminus Y, Y \setminus X)
\]

\[
\leq d_{\bar{\chi}}(Y \setminus X, X \setminus Y)
\]

\[
\leq n_\mathsf{S} + L \cdot \max\{n_1, n_\mathsf{D}\}.
\]

Case 2: \(n_1 > n_\mathsf{D}\). In this case, there exists an injection \(\chi' : X_\mathsf{D} \rightarrow Y_1\) and we can let \(\bar{\chi} : X \setminus Y \rightarrow Y \setminus X\) be such that

\[
\bar{\chi}(x) = \begin{cases} 
\chi(x) & \text{if } x \in X_\mathsf{S}; \\
\chi'(x) & \text{if } x \in X_\mathsf{D}.
\end{cases}
\]

Since \(|Y \setminus X| - |X \setminus Y| = |Y_1| - |X_\mathsf{D}| = n_1 - n_\mathsf{D}\), then by (1),

\[
d_{\bar{\chi}}(X \setminus Y, Y \setminus X) = \sum_{x \in X \setminus Y} d_H(x, \bar{\chi}(x)) + L \cdot (|Y \setminus X| - |X \setminus Y|)
\]

\[
= \sum_{x \in X_\mathsf{S}} d_H(x, \chi(x)) + \sum_{x \in X_\mathsf{D}} d_H(x, \chi'(x)) + L \cdot (n_1 - n_\mathsf{D})
\]

\[
\leq n_\mathsf{S} + L \cdot n_1 + L \cdot (n_1 - n_\mathsf{D})
\]

\[
= n_\mathsf{S} + L \cdot n_1
\]

\[
= n_\mathsf{S} + L \cdot \max\{n_1, n_\mathsf{D}\}.
\]

Hence, similar to Case 1, we have

\[
d_S(X, Y) = d_S(Y \setminus X, X \setminus Y)
\]

\[
\leq d_{\bar{\chi}}(X \setminus Y, Y \setminus X)
\]

\[
\leq n_\mathsf{S} + L \cdot \max\{n_1, n_\mathsf{D}\}.
\]

In both cases, we have \(d_S(X, Y) \leq n_\mathsf{S} + L \cdot \max\{n_1, n_\mathsf{D}\}\), which completes the proof. \(\blacksquare\)
For the decoder, when receiving a subset $Y \subseteq \mathbb{A}_L$, its task is to find a possible input subset $\hat{X} \subseteq \mathbb{A}_L$ that is most similar to $Y$. By the above discussion and Corollary 11, clearly, the sequence-subset distance is a good choice of metric for similarity between $Y$ and $X$. In the next subsection, we will discuss error correction in DNA storage channel using codes with respect to sequence-subset distance.

C. Codes with Sequence-Subset Distance

A sequence-subset code over $\mathbb{A}_L$ is a subset $C$ of the power set $\mathcal{P}(\mathbb{A}_L)$ of the set $\mathbb{A}_L$. We call each element of $\mathbb{A}_L$ a sequence and call $L$ the sequence length of $C$. The size $|C|$ of $C$ is called the code size of $C$. In contrast, for each codeword $X \in C$, the size of $X$ (i.e., the number of sequences contained in $X$) is called the codeword size of $C$. The maximum of codeword sizes of $C$, i.e., $M = \max \{|X|; X \in C\}$, is called the maximal codeword size of $C$. A sequence-subset code $C$ is said to have constant codeword size if all codewords of $C$ have the same codeword size.

The code rate of $C$ is defined as

$$R = \frac{\log_q |C|}{\log_q \left(\sum_{m=0}^{M} \binom{q}{m}\right)},$$

where $q = |\mathbb{A}|$ and $\sum_{m=0}^{M} \binom{q}{m}$ is the number of all subsets of $\mathbb{A}_L$ of size not greater than $M$.

The minimum distance of a sequence-subset code $C$, denoted by $d_S(C)$, is the minimum of the sequence-subset distance between any two distinct codewords of $C$, that is,

$$d_S(C) = \min \{d_S(X, X'); X, X' \in C \text{ and } X \neq X'\}.$$

In general, $L, M, |C|$ and $d_S(C)$ are three main parameters of $C$, and we will call $C$ an $(L, M, |C|, d_S(C))_q$ code, where $q$ is the size of the alphabet $\mathbb{A}$.

Let $C \subseteq \mathcal{P}(\mathbb{A}_L)$ be a sequence-subset code. We denote $\overline{C} = \{\overline{X}; X \in C\}$, where $\overline{X} = \mathbb{A}_L \setminus X$. By Corollary 1 for any $X_1, X_2 \in C$, we have $d_S(X_1, X_2) = d_S(X_1 \setminus X_2, X_2 \setminus X_1) = d_S(\overline{X_1}, \overline{X_2})$. So $\overline{C}$ and $C$ have the same sequence length $L$, code size $|\overline{C}| = |C|$ and minimum distance $d_S(\overline{C}) = d_S(C)$. Hence, for sequence-subset code with constant codeword size $M$, it is reasonable to assume $M \leq \frac{|\mathbb{A}_L|}{2}$. Otherwise, we can consider $\overline{C}$, which has constant codeword size $\overline{M} = |\mathbb{A}_L| - M \leq \frac{|\mathbb{A}_L|}{2}$.

A minimum-distance decoder for $C$ is a function $D: \mathcal{P}(\mathbb{A}_L) \rightarrow C$ such that for any $Y \in \mathcal{P}(\mathbb{A}_L)$,

$$D(Y) = \arg\min_{X' \in C} d_S(X', Y).$$

Theorem 2: Suppose $C$ has minimum distance $d_S(C)$ and

$$n_S + L \cdot \max\{n_1, n_2\} \leq \frac{d_S(C) - 1}{2}. \tag{3}$$

Then any error of pattern $(n_1, n_2, n_3)$ can be corrected by the minimum-distance decoder for $C$.

Proof: Let $X$ be the set of input sequences and $Y$ be the set of output sequences of the DNA storage channel. By Lemma 3 if $Y$ has error pattern $(n_1, n_2, n_3)$, then

$$d_S(X, Y) \leq n_S + L \cdot \max\{n_1, n_2\}.$$

Combining this with (3), we have

$$d_S(X, Y) \leq \frac{d_S(C) - 1}{2}.$$

So $X = \arg\min_{Y \in C} d_S(X', Y) = D(Y)$, and hence $X$ can be correctly recovered by the minimum distance decoder. \hfill \blacksquare

In (11) and (12), it was assumed that the number of output sequences is always smaller than the number of input sequences. In this work, we dismiss this assumption and allow the number of output sequences of the DNA storage channel to be larger than the number of input sequences.

III. BOUNDS ON THE SIZE OF SEQUENCE SUBSET CODES

Let $S_q(L, M, d)$ denote the maximum number of codewords in a sequence-subset code over a $q$-ary alphabet with sequence length $L$, constant codeword size $M$ and minimum sequence-subset distance at least $d$. A $q$-ary sequence-subset code is said to be optimal (with respect to code size) if it has the largest possible code size of any $q$-ary sequence-subset code of the given parameters $L, M$ and $d$. In this section, we always assume that $\mathbb{A}$ is an alphabet of size $q$. We will derive some upper bounds on $S_q(L, M, d)$.

Clearly, for any sequence-subset code $C \subseteq \mathcal{P}(\mathbb{A}_L)$ with constant codeword size $M$, its minimum distance $d_S(C) \leq LM$, and hence $M \geq \frac{2d_S(C)}{L}$. For this reason, in the following, we always assume that $d \leq LM$, or equivalently, $M \geq \frac{d}{L}$. 

A. Upper Bound for the Special Case $L \mid d$

First, consider the special case that $M_0 = \frac{d}{q}$. Since $M_0$ is an integer, we need to further assume that $L \mid d$. Then we have the following upper bound on $S_q(L, M_0, d)$.

**Theorem 3:** Suppose $L \mid d$ and $M_0 = \frac{d}{q}$. We have

$$S_q(L, M_0, d) \leq \left\lfloor qM_0^{\frac{d}{q}} \right\rfloor.$$

**Proof:** Let $C = \{X_1, X_2, \cdots, X_N\} \subseteq \mathcal{P}(\mathbb{A}^L)$ be an arbitrary sequence-subset code with constant codeword size $M_0$ and minimum distance $d$, where for each $i \in [N]$, $X_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,M_0}\} \subseteq \mathbb{A}^L$. We need to prove $N \leq qM_0^{\frac{d}{q}}$. For each $\ell \in [L]$ and $i \in [N]$, let

$$W_{i,\ell} = \bigcup_{j \in [M_0]} \{x_{i,j}(\ell)\}.$$ 

Note that the minimum distance of $C$ is $d = LM_0$. Then from Definition 1 it is necessary that for any distinct $i_1, i_2 \in [N]$ and any (not necessarily distinct) $j_1, j_2 \in [M_0]$, $d_H(x_{i_1,j_1}, x_{i_2,j_2}) = L$, which implies that for any $\ell \in [L]$ and any $(j_1, j_2, \cdots, j_N) \in [M_0]^N$, $x_{i_1,j_1}(\ell), x_{i_2,j_2}(\ell), \cdots, x_{i_N,j_N}(\ell)$ are distinct elements of $\mathbb{A}$. Hence, for each fixed $\ell \in [L]$ and $i \in [N]$, $W_{i,\ell}, W_{2,\ell}, \cdots, W_{N,\ell}$ are mutually disjoint subsets of $\mathbb{A}$, which implies that

$$\sum_{i=1}^{N} |W_{i,\ell}| \leq |\mathbb{A}| = q. \tag{5}$$

By the construction of $W_{i,\ell}$, for each $i \in [N]$ and $j \in [M_0]$, we have $x_{i,j} \in W_{i,1} \times W_{i,2} \times \cdots \times W_{i,L}$, which implies that $X_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,M_0}\} \subseteq W_{i,1} \times W_{i,2} \times \cdots \times W_{i,L}$, and hence we have

$$|W_{i,1} \times W_{i,2} \times \cdots \times W_{i,L}| = \prod_{\ell=1}^{L} |W_{i,\ell}| \geq |X_i| = M_0. \tag{6}$$

Now, consider (5). By the inequality of arithmetic and geometric means, for each $\ell \in [L]$, we have

$$\frac{q}{N} \geq \frac{1}{N} \sum_{i=1}^{N} |W_{i,\ell}| \geq \left( \prod_{i=1}^{N} |W_{i,\ell}| \right)^{\frac{1}{N}}. \tag{7}$$

Combining this with (6), we have

$$\left( \frac{q}{N} \right)^L \geq \prod_{\ell=1}^{L} \left( \prod_{i=1}^{N} |W_{i,\ell}| \right)^{\frac{1}{N}} = \prod_{i=1}^{N} \left( \prod_{\ell=1}^{L} |W_{i,\ell}| \right)^{\frac{1}{N}} \geq (M_0^{\frac{1}{N}})^L = M_0.$$ 

From this we have $\frac{q}{N} \geq M_0^{\frac{d}{q}}$, which implies $N \leq qM_0^{\frac{d}{q}}$. Hence,

$$S_q(L, M_0, d) \leq qM_0^{\frac{d}{q}}.$$ 

Since $S_q(L, M_0, d)$ is an integer, so

$$S_q(L, M_0, d) \leq \left\lfloor qM_0^{\frac{d}{q}} \right\rfloor,$$

which completes the proof. \hfill \blacksquare

B. Plotkin-like Bound

We present the Plotkin-like Bound of sequence-subset codes as the following theorem.

**Theorem 4 (Plotkin-like Bound):** Let $C$ be an $(L, M, N, d)_q$ code such that $rLM < d$, where $r = 1 - \frac{1}{q}$. Then

$$N \leq \frac{d}{d - rLM}.$$ 

**Proof:** Our proof of this theorem is similar to the proof of [17, Theorem 2.2.1].
Suppose $C = \{X_1, X_2, \ldots, X_N\}$ such that for each $i \in [N]$, $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,M}\} \subseteq \mathbb{A}^L$. First, we have the following claim, which we will prove later.

**Claim 1:** For any distinct $i_1, i_2 \in [N]$, we have

$$d_S(X_{i_1}, X_{i_2}) \leq \frac{1}{M} \sum_{j_1, j_2 \in [M]} d_H(x_{i_1,j_1}, x_{i_2,j_2}).$$

Now, let

$$A = \sum_{i_1, i_2 \in [N]} \sum_{j_1, j_2 \in [M]} d_H(x_{i_1,j_1}, x_{i_2,j_2}).$$

Since $d$ is the minimum distance of $C$, by the averaging principle [18], we have

$$d \leq \left(\frac{N}{2}\right)^{-1} \sum_{\{i_1, i_2\} \subseteq [N]} d_S(X_{i_1}, X_{i_2})$$

$$= \frac{1}{2} \left(\frac{N}{2}\right)^{-1} \sum_{i_1, i_2 \in [N], i_1 \neq i_2} d_S(X_{i_1}, X_{i_2})$$

$$\leq \frac{1}{2} \left(\frac{N}{2}\right)^{-1} \sum_{i_1, i_2 \in [N]} d_S(X_{i_1}, X_{i_2})$$

$$\leq \frac{1}{N(N-1)} \sum_{i_1, i_2 \in [N]} \left(\frac{1}{M} \sum_{j_1, j_2 \in [M]} d_H(x_{i_1,j_1}, x_{i_2,j_2})\right)$$

$$= \frac{1}{N(N-1)} \frac{1}{M} \cdot A,$$

where the last inequality is obtained by Claim 1.

For each $a \in \mathbb{A}$ and $\ell \in [L]$, let $n_{\ell,a}$ be the number of $(i, j) \in [N] \times [M]$ such that $x_{i,j}(\ell) = a$. Then for each fixed $\ell \in [L]$, we have

$$\sum_{a \in \mathbb{A}} n_{\ell,a} = NM. \quad (8)$$

Moreover, we have

$$A = \sum_{i_1, i_2 \in [N]} \sum_{j_1, j_2 \in [M]} d_H(x_{i_1,j_1}, x_{i_2,j_2})$$

$$= \sum_{\ell=1}^{L} \sum_{a \in \mathbb{A}} n_{\ell,a} (NM - n_{\ell,a})$$

$$= L(NM)^2 - \sum_{\ell=1}^{L} \sum_{a \in \mathbb{A}} n_{\ell,a}^2. \quad (9)$$

For each $\ell \in [L]$, by the Cauchy-Schwartz inequality,

$$\left(\sum_{a \in \mathbb{A}} n_{\ell,a}\right)^2 \leq q \sum_{a \in \mathbb{A}} n_{\ell,a}^2,$$

where $q = |\mathbb{A}|$. Combining this with (8), we obtain

$$A \leq L(NM)^2 - \sum_{\ell=1}^{L} \frac{1}{q} \left(\sum_{a \in \mathbb{A}} n_{\ell,a}\right)^2$$

$$= L(NM)^2 - \sum_{\ell=1}^{L} \frac{1}{q} (NM)^2$$

$$= \left(1 - \frac{1}{q}\right) L(NM)^2,$$

where the first equality is obtained from (8). Combining (7) and (10), we obtain

$$d \leq \frac{1}{N(N-1)} \frac{1}{M} \left(1 - \frac{1}{q}\right) L(NM)^2.$$
Solving $N$ from the above inequality we obtain

$$N \leq \frac{d}{d - rLM},$$

where $r = 1 - \frac{1}{q}$.

To complete the proof of Theorem 4 we still need to prove Claim 1.

**Proof of Claim 1:** Let $\mathcal{S}_M$ denote the permutation group on $[M]$. Note that for any $j_1, j_2 \in [M]$, not necessarily distinct, there are $(M - 1)!$ permutations $\chi \in \mathcal{S}_M$ such that $\chi(j_1) = j_2$. So we have

$$\sum_{\chi \in \mathcal{S}_M} \sum_{j \in [M]} d_H(x_{i_1,j}, x_{i_2,\chi(j)}) = (M - 1)! \sum_{j_1, j_2 \in [M]} d_H(x_{i_1,j_1}, x_{i_2,j_2}).$$

(11)

Further, by Definition [1] and the averaging principle [18], we have

$$d_S(x_{i_1}, x_{i_2}) \leq \frac{1}{M!} \sum_{\chi \in \mathcal{S}_M} d_H(x_{i_1}, x_{i_2})$$

$$= \frac{1}{M!} \sum_{\chi \in \mathcal{S}_M} \sum_{j \in [M]} d_H(x_{i_1,j}, x_{i_2,\chi(j)})$$

$$= (M - 1)! \sum_{j_1, j_2 \in [M]} d_H(x_{i_1,j_1}, x_{i_2,j_2})$$

$$= \frac{1}{M} \sum_{j_1, j_2 \in [M]} d_H(x_{i_1,j_1}, x_{i_2,j_2}),$$

where the second equality comes from (11).

**C. Singleton-like Bound**

For each code $C = \{X_1, X_2, \ldots, X_N\} \subseteq \mathcal{P}(\mathbb{A}^L)$, denote

$$V(C) = \bigcup_{i=1}^{N} X_i.$$  

(12)

Further, let $\bar{S}_q(L, M, K, d)$ denote the maximum number of codewords in a sequence-subset code $C$ over a $q$-ary alphabet $\mathbb{A}$ with sequence length $L$, constant codeword size $M$, minimum sequence-subset distance at least $d$ and $|V(C)| \leq K$. Clearly, for any $K \leq q^L$,

$$\bar{S}_q(L, M, K, d) \leq \bar{S}_q(L, M, q^L, d) = S_q(L, M, d).$$

(13)

We first prove a recursive bound on $\bar{S}_q(L, M, K, d)$ as the following theorem.

**Theorem 5:** Suppose $d \leq LM$ and $K \leq q^L$. We have

$$\bar{S}_q(L, M, K, d) \leq \frac{K}{M} \bar{S}_q(L, M - 1, K - 1, d).$$

(14)

**Proof:** Let $C = \{X_1, X_2, \ldots, X_N\} \subseteq \mathcal{P}(\mathbb{A}^L)$ be a sequence-subset code with constant codeword size $M$, minimum distance at least $d$ such that $|V(C)| \leq K$ and code size $|C| = N = \bar{S}_q(L, M, K, d)$, where $X_i \subseteq \mathbb{A}^L$ for each $i \in [N]$.

For each $x \in V(C)$, let

$$C(x) = \{X \in C; x \in X\}$$

and

$$\overline{C}(x) = \{X = X \setminus \{x\}; X \in C(x)\}.$$  

Then $\overline{C}(x) \subseteq \mathcal{P}(\mathbb{A}^L)$ has constant codeword size $M - 1$, size $|\overline{C}(x)| = |C(x)|$ and $|V(\overline{C}(x))| \leq K - 1$.

Moreover, for any distinct $X_{i_1}, X_{i_2} \in C(x)$, by the construction of $\overline{C}(x)$, we have $X_{i_1} = X_{i_1 \setminus \{x\}}$ and $X_{i_2} = X_{i_2 \setminus \{x\}}$ for some distinct $X_{i_1}, X_{i_2} \in C(x)$. So $X_{i_1 \setminus X_{i_2}} = X_{i_1} \setminus X_{i_2}$ and $X_{i_2 \setminus X_{i_1}} = X_{i_2} \setminus X_{i_1}$, and hence by Corollary [1]

$$d_S(X_{i_1}, X_{i_2}) = d_S(X_{i_1 \setminus X_{i_2}}, X_{i_2 \setminus X_{i_1}}).$$

So we have $d_S(\overline{C}(x)) = d_S(C(x))$. On the other hand, since $C(x) \subseteq C$, then $d_S(C(x)) \geq d_S(C) \geq d$. Hence, $d_S(\overline{C}(x)) \geq d$. 


Combining this with (17), we have

\[ |C(x)| \leq S_q(L, M - 1, K - 1, d). \] (15)

Now, we estimate \(|C(x)|\). Since \(|C(x)| = |C(x)|\), it is sufficient to estimate \(|C(x)|\). Denote \(V(C) = \{x_1, x_2, \ldots, x_K\}\), where \(K = |V(C)|\). Consider the \(N \times K\) matrix \(I = (a_{i,j})\) such that \(a_{i,j} = 1\) if \(x_j \in X_i\), and \(a_{i,j} = 0\) otherwise. Note that the number of ones in row \(i\) of \(I\) is \(|X_i| = M\) and the number of ones in column \(j\) of \(I\) is \(|C(x_j)|\). By counting the total number of ones in \(I\), we obtain

\[
\sum_{x \in V(C)} |C(x)| = \sum_{x \in C} |x| = MN.
\]

By the averaging principle [18], there exists an \(x_{j_0} \in V(C)\) such that

\[
|C(x_{j_0})| \geq \frac{MN}{|V(C)|} = \frac{MN}{K}.
\]

Hence,

\[
N \leq \frac{K}{M} |C(x_{j_0})| = \frac{K}{M} |\overline{C}(x_{j_0})|.
\]

Note that \(|C| = \overline{S}_q(L, M, K, d) = N\). Then we have

\[
\overline{S}_q(L, M, K, d) \leq \frac{K}{M} |\overline{C}(x_0)|.
\]

This, combining with (15), implies that

\[
\overline{S}_q(L, M, K, d) \leq \frac{K}{M} \overline{S}_q(L, M - 1, K - 1, d),
\]

which completes the proof.

Now, we can prove a Singleton-like bound for sequence-subset codes as follows.

**Theorem 6 (Singleton-like Bound):** Suppose \(rLM_0 < d \leq LM_0\), where \(r = 1 - \frac{1}{q}\) and \(M_0 = \left\lfloor \frac{M}{q} \right\rfloor\). Then

\[
S_q(L, M, d) \leq \left( \frac{M - M_0 - 1}{M - k} \right) \cdot f(L, M_0, q),
\]

where

\[
f(L, M_0, q) = \begin{cases} 
qM_0^{-r} & \text{if } d = LM_0; \\
\frac{d}{d - rLM_0} & \text{if } rLM_0 < d < LM_0.
\end{cases}
\] (16)

**Proof:** Denote \(\overline{M} = M - M_0\). Repeatedly using Theorem 5, we obtain

\[
\overline{S}_q(L, M, q^L, d) \leq \left( \prod_{k=0}^{M-1} \frac{q^L - k}{\overline{M} - k} \right) \overline{S}_q(L, M_0, q^L - \overline{M}, d).
\]

Moreover, according to (13), we have

\[
S_q(L, M, d) = \overline{S}_q(L, M, q^L, d)
\]

and

\[
\overline{S}_q(L, M_0, q^L - \overline{M}, d) \leq S_q(L, M_0, q^L, d) = S_q(L, M_0, d).
\]

Combining the above three equations, we have

\[
S_q(L, M, d) \leq \left( \prod_{k=0}^{M-M_0-1} \frac{q^L - k}{\overline{M} - k} \right) \cdot S_q(L, M_0, d).
\] (17)

Let \(f(L, M_0, q)\) be defined as in (16). By Theorem 5 and Theorem 4, we have

\[
S_q(L, M_0, d) \leq f(L, M_0, q).
\]

Combining this with (17), we have

\[
S_q(L, M, d) \leq \left( \prod_{k=0}^{M-M_0-1} \frac{q^L - k}{\overline{M} - k} \right) \cdot f(L, M_0, q),
\]
which completes the proof.

Remark 1: It is easy to see that

$$\left(\frac{q^L}{M}\right) = \left(\prod_{k=0}^{M-M_0-1} \frac{q^L - k}{M - k}\right) \left(\frac{q^L - M + M_0}{M_0}\right).$$

So the bound in Theorem 6 gives a bound on the code rate as

$$\frac{S_q(L, M, d)}{\left(\frac{q^L}{M}\right)} \leq \frac{1}{\left(\frac{q^L - M + M_0}{M_0}\right)} \cdot f(L, M_0, q),$$

where $f(L, M_0, q)$ is defined as in (16).

IV. CONSTRUCTIONS OF SEQUENCE-SUBSET CODES

In this section, we give some constructions of sequence-subset codes. As in Section III, we will always assume that $\mathcal{A}$ is an alphabet of size $q$.

A. Construction of Optimal Codes

In this subsection, we give a construction of optimal $(L, M_0, d)_q$ code (with respect to code size) for the special case that $L | d$ and $M_0^\frac{1}{d}$ is an integer, where $M_0 = \frac{d}{L}$.

Theorem 7: Suppose $L | d$ and $M_0^\frac{1}{d}$ is an integer, where $M_0 = \frac{d}{L}$. There exists an $(L, M_0, d)_q$ sequence-subset code whose code size is $N = \left[qM_0^{-\frac{1}{d}}\right]$.  

Proof: Since $N = \left[qM_0^{-\frac{1}{d}}\right]$, we have $N \leq qM_0^{-\frac{1}{d}}$, and hence

$$q \geq NM_0^\frac{1}{d}.$$ 

So we can partition $\mathcal{A}$ into $N$ mutually disjoint subsets $W_1, W_1, \cdots, W_N$ such that for each $i \in [N]$, $|W_i| \geq M_0^\frac{1}{d}$. So the size of the Cartesian product $W_i^L$ of $L$ copies of $W_i$ is greater than $|M_0|$, and hence we can pick a subset $X_i = \{x_1, x_2, \cdots, x_l, M_0\} \subseteq W_i^L$. Now, let $\mathcal{C} = \{X_i; i \in [N]\}$. Then $\mathcal{C} \subseteq \mathcal{P}(\mathcal{A}^L)$ is a sequence-subset code with constant codeword size $M_0$ and $|\mathcal{C}| = N = \left[qM_0^{-\frac{1}{d}}\right]$. Moreover, since $W_1, W_1, \cdots, W_N$ are mutually disjoint, it is easy to verify that for any distinct $i_1, i_2 \in [N]$ and any $j_1, j_2 \in [M_0]$,

$$d_H(x_{i_1,j_1}, x_{i_2,j_2}) = L.$$ 

So for any distinct $i_1, i_2 \in [N]$,

$$d_S(x_{i_1}, x_{i_2}) = LM_0 = d,$$

which implies that $d_S(C) = d$. 

In summary, $\mathcal{C}$ is an $(L, M_0, d)_q$ sequence-subset code of size $N = \left[qM_0^{-\frac{1}{d}}\right]$. 

Note that by Theorem 3 if $L | d$ and $M_0 = \frac{d}{L}$, then $S_q(L, M_0, d) \leq \left[qM_0^{-\frac{1}{d}}\right]$. So the code $\mathcal{C}$ constructed in Theorem 7 is optimal with respect to code size, and we have the following corollary.

Corollary 2: Suppose $L | d$ and $M_0^\frac{1}{d}$ is an integer, where $M_0 = \frac{d}{L}$. We have

$$S_q(L, M_0, d) = \left[qM_0^{-\frac{1}{d}}\right].$$

B. Construction Based on Binary Codes

In the rest of this section, to distinguish from sequence-subset code (i.e., a subset of the power set $\mathcal{P}(\mathcal{A}^L)$ of the set $\mathcal{A}^L$), we will call any subset of $\mathcal{A}^L$ a conventional code. An $(L, N, d)_q$ conventional code is a subset of $\mathcal{A}^L$ with $N$ codewords and minimum Hamming distance $d$ (recalling that $q$ is the size of the alphabet $\mathcal{A}$). Our following constructions of sequence-subset codes are based on conventional codes with respect to Hamming distance.

The construction given in this subsection is a modification of the Construction 2 of (11).

Let $C_1 = \{x_1, x_2, \cdots, x_K\} \subseteq \mathcal{A}^L$ be a conventional code over $\mathcal{A}$ and $C_2 = \{w_1, w_2, \cdots, w_N\} \subseteq \mathcal{P}^L$ be a conventional binary code. For each $w_i \in C_2$, let $X_i = \{x_j; j \in \text{supp}(w_i)\}$, where $\text{supp}(w_i) = \{j \in [K]; w_i(j) \neq 0\}$ is the support of $w_i$. Further, let $C = \{X_1, X_2, \cdots, X_N\}$. 

Moreover, since \( \cdot \) is an arbitrary injection. So by Definition 1 and Corollary 1, we have

\[
d_{\chi}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \sum_{x \in \mathbf{x}_{i_1}} d_{H}(x, \chi(x)) + L(|\mathbf{x}_{i_2}|-|\mathbf{x}_{i_1}|).
\]

(18)

Since \( C_1 \) has minimum (Hamming) distance \( d_1 \) and by construction of \( C \), \( x \) and \( \chi(x) \) are distinct codeword in \( C_1 \), so

\[
\sum_{x \in \mathbf{x}_{i_1}} d_{H}(x, \chi(x)) \geq |\mathbf{x}_{i_1}| \cdot d_1.
\]

Moreover, since \( C_1 \subseteq \mathbb{A}^L \), then \( L \geq d_1 \). Hence, (18) implies that

\[
d_{\chi}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) \geq |\mathbf{x}_{i_1}| \cdot d_1 + d_1(|\mathbf{x}_{i_2}| - |\mathbf{x}_{i_1}|)
= d_1 \cdot |\mathbf{x}_{i_2}|
= d_1 \cdot |\mathbf{x}_{i_2} \setminus \mathbf{x}_{i_1}|.
\]

(19)

By the construction of \( C \), \( \mathbf{x}_{i_1} = \{ x_j : j \in \text{supp}(w_{i_1}) \} \) and \( \mathbf{x}_{i_2} = \{ x_j : j \in \text{supp}(w_{i_2}) \} \) for some distinct \( w_{i_1}, w_{i_2} \in C_2 \). Then we have

\[
|\mathbf{x}_{i_1} \setminus \mathbf{x}_{i_2}| + |\mathbf{x}_{i_2} \setminus \mathbf{x}_{i_1}| = d_{H}(w_{i_1}, w_{i_2}) \geq d_2,
\]

where \( d_2 \) is the minimum (Hamming) distance of \( C_2 \). Note that \( |\mathbf{x}_{i_1} \setminus \mathbf{x}_{i_2}| \leq |\mathbf{x}_{i_2} \setminus \mathbf{x}_{i_1}| \). Then by the above equation, we have

\[
|\mathbf{x}_{i_2} \setminus \mathbf{x}_{i_1}| \geq \left\lceil \frac{d_2}{2} \right\rceil.
\]

Combining this with (19), we have

\[
d_{\chi}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) \geq d_1 \cdot \frac{d_2}{2}.
\]

Note that \( \chi : \mathbf{x}_{i_1} \setminus \mathbf{x}_{i_2} \to \mathbf{x}_{i_2} \setminus \mathbf{x}_{i_1} \) is an arbitrary injection. So by Definition 1 and Corollary 1 we have

\[
d_S(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = d_S(\mathbf{x}_{i_1} \setminus \mathbf{x}_{i_2}, \mathbf{x}_{i_2} \setminus \mathbf{x}_{i_1}) \geq d_1 \cdot \left\lfloor \frac{d_2}{2} \right\rfloor,
\]

which completes the proof. \( \square \)

**Remark 2:** The code \( C \) constructed in this subsection may or may not have constant codeword size, depending on whether \( C_2 \) is a constant weight binary code. In fact, if \( C_2 \) is a constant weight code, then \( C \) has constant codeword size. Otherwise, \( C \) does not have constant codeword size.

**C. Construction Based on Non-binary Codes**

Let \( \mathbb{A} \) and \( \mathbb{B} \) be two alphabets of size \( q \) and \( \tilde{q} \), respectively. Let \( C_1 \) be an \((L, M\tilde{q}, d_1)_{q} \) conventional code over \( \mathbb{A} \) and \( C_2 \) be an \((M, N, d_2)_{\tilde{q}} \) conventional code over \( \mathbb{B} \). The \( M\tilde{q} \) codewords of \( C_1 \) can be indexed as

\[
C_1 = \{ x_{i,j} : i \in [M], j \in \mathbb{B} \}.
\]

Then from each \( \mathbf{c} = (c_1, c_2, \ldots, c_M) \in C_2 \), we can obtain a subset

\[
\mathbf{X}_\mathbf{c} = \{ \mathbf{x}_{1,c_1}, \mathbf{x}_{2,c_2}, \ldots, \mathbf{x}_{M,c_M} \} \subseteq C_1.
\]

Let

\[
C = \{ \mathbf{X}_\mathbf{c} : \mathbf{c} \in C_2 \}.
\]

(20)
Then $C$ is a sequence-subset code over $A$ and we have the following theorem.

**Theorem 9:** The code $C$ constructed by (20) has sequence length $L$, constant codeword size $M$, code size $|C| = N$, and minimum sequence-subset distance

$$d_S(C) \geq d_1d_2.$$  

**Proof:** From the construction it is easy to see that $C$ has sequence length $L$, constant codeword size $M$ and code size $|C| = N$. It remains to prove that $d_S(C) \geq d_1d_2$, that is, $d_S(X_c, X_{c'}) \geq d_1d_2$ for any distinct $X_c$ and $X_{c'}$ in $C$, where $c = (c_1, c_2, \ldots, c_M)$ and $c' = (c'_1, c'_2, \ldots, c'_M)$ are any pair of distinct codewords in $C_2$.

Let $A$ be the set of all $i \in [M]$ such that $c_i \neq c'_i$. Since $C_2$ has minimum (Hamming) distance $d_2$, then

$$|A| = d_H(c, c') \geq d_2.$$

Denote

$${X}_c = \{x_{i,c_i}; i \in A\} \quad \text{and} \quad {X}_c' = \{x_{i,c'_i}; i \in A\}.$$  

Then by the construction, we have

$${X}_c = X_c \setminus X_{c'} \quad \text{and} \quad {X}_c' = X_{c'} \setminus X_c.$$  

So by Corollary 1, it suffices to prove that $d_S({X}_c, {X}_c') \geq d_1d_2$.

Note that $|{X}_c| = |{X}_c'| = |A|$ and $X_c \cap X_c' = \emptyset$. Then for any injection $\chi : X_c \to X_c'$, we have

$$d_{\chi}(X_c, X_c') = \sum_{x \in X_c} d_H(x, \chi(x)) \geq |A| \cdot d_1 \geq d_1d_2,$$

where the equality comes from (28), the first inequality comes from the assumption that $C_1$ has minimum (Hamming) distance $d_1$, and the second inequality comes from the fact that $|A| \geq d_2$. By Definition 28 $d_S(X_c, X_{c'}) \geq d_1d_2$, and hence by Corollary 1 $d_S(X_c, X_{c'}) \geq d_1d_2$. Since $X_c$ and $X_{c'}$ are any pair of distinct codewords in $C$, we have $d_S(C) \geq d_1d_2$, which completes the proof.

The following example is a special case of this construction.

**Example 1:** Let $C_1$ be an $[L, k, d_1]_{q'}$ linear code such that the first $k$ symbols of the codewords of $C_1$ are the information symbols. For any given integer $r$ such that $1 \leq r < k$, let $q = q^r$ and $M = q^r$, where $s = k - r$. Note that there exists a bijection $\pi : [M] \to \mathbb{F}_{q^r}^r$. Moreover, fixing a basis, each element of $\mathbb{F}_{q^r}$ can be uniquely represented as a vector in $\mathbb{F}_{q^r}^r$, so we can identify each element of $\mathbb{F}_{q^r}$ as a vector in $\mathbb{F}_{q^r}^r$. Then for each $i \in [M]$ and each $j \in \mathbb{F}_{q^r}$, we can let

$$x_{i,j} = (x_1, x_2, \ldots, x_L);$$

$$(x_1, x_2, \ldots, x_s) = \pi(i) \quad \text{and} \quad (x_{s+1}, \ldots, x_k) = j.$$

Now, let $C_2$ be an $[M, K, d_2]_{q^r}$ linear code, where $K \in [M]$ is another design parameter. Then for each $c = (c_1, c_2, \ldots, c_M) \in C_2$, we can obtain

$$X_c = \{x_{1,c_1}, x_{2,c_2}, \ldots, x_{M,c_M}\} \subseteq C_1,$$

that is, for each $i \in [M]$, $x_{i,c_i} = (x_1, x_2, \ldots, x_L)$ such that

$$(x_1, x_2, \ldots, x_s) = \pi(i) \quad \text{and} \quad (x_{s+1}, \ldots, x_k) = c_i.$$  

Finally, we have

$$C = \{X_c; c \in C_2\}.$$  

The construction method of this special case is essentially similar to the method used in (28).

**D. Construction Based on Sequence Index**

In this subsection, if $x = (x(1), x(2), \ldots, x(L)) \in A^L$ and $I = \{i_1, i_2, \ldots, i_m\} \subseteq [L]$ such that $i_1 < i_2 < \cdots < i_m$, then we denote $x(I) = (x(i_1), x(i_2), \ldots, x(i_m))$.

The construction given in this subsection is an improvement of the Construction 1 of (30).

Let $C_1 = \{s_1, s_2, \ldots, s_M\} \subseteq A^{L_1}$ be a conventional code over $A$ with block length $L_1$ and minimum (Hamming) distance $d_1$, and $C_2 = \{u_1, u_2, \ldots, u_N\} \subseteq A^{d_1 M}$ be a conventional code over $A$ with block length $d_1 M$ and minimum (Hamming) distance $d_2$. For each $j \in [M]$, let

$$I_j = \{\ell \in \mathbb{Z}; (j-1)d_1 < \ell \leq jd_1\}$$
and for each $i \in [N]$, let

$$\mathbf{X}_i = \{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,M}\}$$

such that for each $j \in [M]$,

$$\mathbf{x}_{i,j} = (s_{j,i}, u_i(I_j)).$$

Finally, let

$$\mathcal{C} = \{\mathbf{X}_i; i \in [N]\}.$$  \hspace{1cm} (21)

Then $\mathcal{C}$ is a sequence-subset code over $\mathcal{A}$. In this construction, each codeword $s_j$ of $C_1$ serves as an index of the sequence $x_{i,j}$ of the codeword $\mathbf{X}_i$, and $u_i(I_j)$ is the information part of $x_{i,j}$. It is the reason that we say this construction is based on sequence index. Moreover, we have the following theorem.

**Theorem 10:** The code $\mathcal{C}$ constructed by (21) has sequence length $L = L_1 + d_1$, constant codeword size $M$, code size $|\mathcal{C}| = N$, and minimum sequence-subset distance

$$d_{S}(\mathcal{C}) \geq d_2.$$  

**Proof:** Clearly, $\mathcal{C}$ has sequence length $L = L_1 + d_1$, constant codeword size $M$ and code size $|\mathcal{C}| = N$. It remains to prove that $d_{S}(\mathcal{C}) \geq d_2$.

Let $i_1, i_2 \in [N]$ be any two distinct elements of $[N]$, we need to prove that $d_{S}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) \geq d_2$, where $\mathbf{X}_{i_1} = \{\mathbf{x}_{i_1,1}, \mathbf{x}_{i_1,2}, \ldots, \mathbf{x}_{i_1,M}\}$ and $\mathbf{X}_{i_2} = \{\mathbf{x}_{i_2,1}, \mathbf{x}_{i_2,2}, \ldots, \mathbf{x}_{i_2,M}\}$. For any permutation $\chi : [M] \rightarrow [M]$, let

$$\mathcal{N} = \{j \in [M]; \chi(j) = j\}$$

and

$$\overline{\mathcal{N}} = \{j \in [M]; \chi(j) \neq j\}.$$  

Then $\mathcal{N} \cap \overline{\mathcal{N}} = \emptyset$ and $\mathcal{N} \cup \overline{\mathcal{N}} = [M]$. Moreover, by (1), we have

$$d_{\chi}(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) = \sum_{j=1}^{M} d_H(\mathbf{x}_{i_1,j}, \mathbf{x}_{i_2,\chi(j)})$$

$$= \sum_{j \in \mathcal{N}} d_H(\mathbf{x}_{i_1,j}, \mathbf{x}_{i_2,\chi(j)}) + \sum_{j \in \overline{\mathcal{N}}} d_H(\mathbf{x}_{i_1,j}, \mathbf{x}_{i_2,\chi(j)})$$

$$= \sum_{j \in \mathcal{N}} d_H(\mathbf{x}_{i_1,j}, \mathbf{x}_{i_2,j}) + \sum_{j \in \overline{\mathcal{N}}} d_H(\mathbf{x}_{i_1,j}, \mathbf{x}_{i_2,j}).$$  \hspace{1cm} (22)

We will estimate the two terms of the right side of Equation (22) separately.

First, by the construction, we have

$$\sum_{j=1}^{M} d_H(\mathbf{x}_{i_1,j}, \mathbf{x}_{i_2,j}) = \sum_{j=1}^{M} d_H(\mathbf{u}_{i_1}(I_j), \mathbf{u}_{i_2}(I_j))$$

$$= d_H(\mathbf{u}_{i_1}, \mathbf{u}_{i_2}) = d_2.$$  

Moreover, since for each $i \in [N]$ and $j \in [M]$, $u_i(I_j)$ has length $d_1$, then again by construction of $\mathcal{C}$, we have

$$d_H(\mathbf{x}_{i_1,j}, \mathbf{x}_{i_2,j}) = d_H(\mathbf{u}_{i_1}(I_j), \mathbf{u}_{i_2}(I_j)) \leq d_1.$$  

Hence, we obtain

$$\sum_{j \in \mathcal{N}} d_H(\mathbf{x}_{i_1,j}, \mathbf{x}_{i_2,j}) = \sum_{j=1}^{M} d_H(\mathbf{x}_{i_1,j}, \mathbf{x}_{i_2,j}) - \sum_{j \in \overline{\mathcal{N}}} d_H(\mathbf{x}_{i_1,j}, \mathbf{x}_{i_2,j})$$

$$= \sum_{j=1}^{M} d_H(\mathbf{u}_{i_1}(I_j), \mathbf{u}_{i_2}(I_j))$$

$$- \sum_{j \in \overline{\mathcal{N}}} d_H(\mathbf{u}_{i_1}(I_j), \mathbf{u}_{i_2}(I_j))$$

$$\geq d_2 - |\overline{\mathcal{N}}| \cdot d_1.$$  

Note that any bijection between $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$ can be uniquely represented by a permutation on the index set $[M]$. So when applying (3) to the pair $(\mathbf{X}_{i_1}, \mathbf{X}_{i_2})$, we can use permutations on $[M]$ to replace bijections between $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$.  


Second, since $C_1$ has minimum (Hamming) distance $d_1$, then by construction of $C$, we have

$$\sum_{j \in \mathcal{N}} d_H(x_{i,j}, x_{i',j'}) \geq \sum_{j \in \mathcal{N}} d_H(s_j, s_{\chi(j)}) \geq |\mathcal{N}| \cdot d_1.$$ 

Combining the above two inequalities with (22), we obtain

$$d_\chi(X_{i_1}, X_{i_2}) = \sum_{j \in \mathcal{N}} d_H(x_{i_1,j}, x_{i_2,j}) + \sum_{j \in \mathcal{N}} d_H(x_{i_1,j}, x_{i_2,\chi(j)}) \geq (d_2 - |\mathcal{N}| \cdot d_1) + |\mathcal{N}| \cdot d_1$$

$$\geq d_2.$$

Since $\chi : [M] \to [M]$ is an arbitrary bijection, then by Definition 1, we have

$$d_S(X_{i_1}, X_{i_2}) \geq d_2.$$

Moreover, since $i_1$ and $i_2$ are any two distinct elements of $[N]$, so we have

$$d_S(C) \geq d_2,$$

which completes the proof.

Remark 3: Using the product of multiple copies of $C$, the construction in this subsection can be further extended as follows. Let $n$ be a given positive integer. For each $n$-tuple $i = (i_1, i_2, \ldots, i_n) \in [N]^n$, let

$$X_i = \{ x_{i,j_1}, x_{i,j_2}, \ldots, x_{i_M} \}$$

such that for each $j \in [M]$, $x_{i,j} = (s_j, u_{i_1}(I_j), \ldots, u_{i_n}(I_j))$.

Finally, let

$$\mathcal{C} = \{ X_i : i = (i_1, i_2, \ldots, i_n) \in [N]^n \}.$$ 

Then the code $\mathcal{C}$ has sequence length $L = L_1 + nd_1$, constant codeword size $M$, code size $|C| = N^n$, and minimum sequence-subset distance

$$d_S(C) \geq d_2.$$ 

V. CONCLUSIONS

We introduced a new metric over the power set of the set of all vectors over a finite alphabet, which generalizes the classical Hamming distance and was used to establish a uniform framework to design error-correcting codes for DNA storage channel. Some upper bounds on the size of the sequence-subset codes were derived and some constructions of such codes were proposed.

It is still an open problem to analyze the tight upper bounds on the size of sequence-subset codes and design optimal codes for general parameters of sequence length, codeword size and minimum distance.

Another interesting problem is how to design sequence-subset codes for DNA storage channel that can be efficiently encoded and decoded.

The sequence-subset distance (Definition 1) can be directly applied to multisets of $\mathbb{A}^L$. So studying the properties of codes over the space of all multisets of $\mathbb{A}^L$ with sequence-subset distance is also a possible research direction.

APPENDIX A

PROOF OF LEMMA 1

If $X_1 \cap X_2 = \emptyset$, the claim is naturally true. So we assume that $X_1 \cap X_2 \neq \emptyset$.

First, we claim that for each $\chi \in \mathcal{X}$ such that $d_S(X_1, X_2) = d_S(X_1, X_2)$ and each $y \in X_1 \cap X_2$, there exists an $x \in X_1$ such that $y = \chi(x)$. This can be proved, by contradiction, as follows. Suppose there is a $y \in X_1 \cap X_2$ such that $y \neq \chi(x')$ for all $x' \in X_1$. Since $y \in X_1 \cap X_2$, then we have $\chi(y) \neq y$, and hence we can let $\chi' : X_1 \to X_2$ be such that $\chi'(y) = y$ and $\chi'(x') = \chi(x')$ for all $x' \in X_1 \setminus \{ y \}$ (see Fig. 2 for an illustration). Note that $d_H(y, \chi(y)) = 0 < d_H(y, \chi'(y))$ and $d_H(x', \chi'(x')) = d_H(x', \chi(x'))$ for all $x' \in X_1 \setminus \{ y \}$. So by (1), we have $d_S(X_1, X_2) < d_S(X_1, X_2)$, which contradicts to (2). Hence, by contradiction, for each $y \in X_1 \cap X_2$, there exists an $x \in X_1$ such that $y = \chi(x)$.

Now, pick a $\chi \in \mathcal{X}$ such that $d_S(X_1, X_2) = d_S(X_1, X_2)$ and denote

$$\mathcal{N}(\chi) = \{ y' \in X_1 \cap X_2 : \chi(y') \neq y' \}.$$ 

If $\mathcal{N}(\chi) = \emptyset$, then by the definition of $\mathcal{N}(\chi)$, $\chi(x) = x$ for all $x \in X_1 \cap X_2$ and we can choose $\chi_0 = \chi$. Otherwise, pick a $y \in \mathcal{N}(\chi)$ and we have $\chi(y) = y_0$ for some $y_0 \in X_2 \setminus \{ y \}$. Moreover, by previous discussion, there exists an $x \in X_1$ such that
Again by construction of $\chi$. So by (2), we have $\chi(y) = y_0$. Then we can modify the injection $\chi$ to a different injection $\chi'$ by letting $\chi'(y) = y$, and the image of all other elements of $X_1$ keep unchanged.

Fig 2. An illustration of the injections in the proof of Lemma 1. For the injection $\chi$, we have $\chi(x) = y$ and $\chi(y) = y_0$, where $y \in X_1 \cap X_2$. We modify the injection $\chi$ to a different injection $\chi'$ by letting $\chi'(x) = y_0$ and $\chi'(y) = y$, and the image of all other elements of $X_1$ keep unchanged.

Fig 3. An illustration of the bijections in the proof of Lemma 1. For the bijection $\chi$, we have $\chi(x) = y$ and $\chi(y) = y_0$, where $y \in X_1 \cap X_2$. We modify the bijection $\chi$ to a different bijection $\chi'$ by letting $\chi'(x) = y_0$ and $\chi'(y) = y$, and the image of all other elements of $X_1$ keep unchanged.

$y = \chi(x)$. Then we can let $\chi': X_1 \to X_2$ be such that $\chi'(x) = y_0$, $\chi'(y) = y$ and $\chi'(x') = \chi_0(x')$ for all $x' \in X_1 \setminus \{x, y\}$ (see Fig. 3 for an illustration). Note that

$$d_H(x, \chi'(x)) + d_H(y, \chi'(y)) = d_H(x, y_0) + d_H(y, y)$$

$$= d_H(x, y_0)$$

$$\leq d_H(x, y) + d_H(y, y_0)$$

$$= d_H(x, \chi(x)) + d_H(y, \chi(y))$$

and by construction of $\chi'$,

$$d_H(x', \chi'(x')) = d_H(x', \chi(x')), \forall x' \in X_1 \setminus \{x, y\}.$$

So by (2), we have

$$d_S(X_1, X_2) = d_S(X_1, X_2) = d_S(X_1, X_2).$$

Again by construction of $\chi'$, we have $\mathcal{N}(\chi') = \mathcal{N}(\chi) \setminus \{y\}$, and hence

$$|\mathcal{N}(\chi')| = |\mathcal{N}(\chi)| - 1,$$

where

$$\mathcal{N}(\chi') = \{y \in X_1 \cap X_2; \chi'(y) \neq y\}.$$ 

If $\mathcal{N}(\chi') = \emptyset$, then $\chi'(x) = x$ for all $x \in X_1 \cap X_2$ and we can choose $\chi_0 = \chi'$. Otherwise, by the same discussion, we can obtain a $\chi'' : X_1 \to X_2$ such that $d_S(X_1, X_2) = d_S(X_1, X_2)$ and $|\mathcal{N}(\chi'')| = |\mathcal{N}(\chi')| - 1$, and so on. Noting that $\mathcal{N}(\chi'') \subseteq X_1 \cap X_2$ is a finite set, we can always find a $\chi_0 \in \mathcal{B}$ such that $d_S(X_1, X_2) = d_S(X_1, X_2)$ and

$$\mathcal{N}(\chi_0) = \{y \in X_1 \cap X_2; \chi_0(x) \neq y\} = \emptyset.$$ 

Hence, we have $\chi_0(x) = x$ for all $x \in X_1 \cap X_2$, which completes the proof.

**APPENDIX B**

**PROOF OF LEMMA 2**

It suffices to prove that if $X'_2 \subseteq X_2$ and $|X_1| \leq |X'_2| = |X_2| - 1$, then

$$d_S(X_1, X'_2) \leq d_S(X_1, X_2).$$ 

Without loss of generality, we can assume

$$X_1 = \{x_1, \ldots, x_n\},$$

$$X'_2 = \{y_1, \ldots, y_{n+1}, \ldots, y_{n+s-1}\}.$$
By Definition 1, we can suppose for any three subsets following three cases.

\[ d_S(X_1, X_2) = \sum_{i=1}^{n} d_H(x_i, y_i) + L(s - 1). \]

By Definition 1 we can suppose
\[ d_S(X_1, X_2) = \sum_{i=1}^{n} d_H(x_i, y_{\ell_i}) + Ls, \]
where \( \{\ell_i; i = 1, 2, \ldots, n\} \) is a subset of \( \{1, 2, \ldots, n + s\} \). We have the following two cases.

Case 1: \( n + s \notin \{\ell_1, \ell_2, \ldots, \ell_n\} \). In this case, we have
\[
d_S(X_1, X_2') = \sum_{i=1}^{n} d_H(x_i, y_{\ell_i}) + L(s - 1) \leq \sum_{i=1}^{n} d_H(x_i, y_{\ell_i}) + L(s - 1) \leq \sum_{i=1}^{n} d_H(x_i, y_{\ell_i}) + Ls = d_S(X_1, X_2),
\]
where the first inequality is obtained by (2).

Case 2: There exists a \( k \in \{1, 2, \ldots, n\} \) such that \( n + s = \ell_k \). Noticing that \( s \geq 1 \), then there exists an \( m \in \{1, 2, \ldots, n + s - 1\} \) such that \( m \notin \{\ell_1, \ell_2, \ldots, \ell_n\} \). Denote \( \ell'_k = m \) and \( \ell'_i = \ell_i \) for \( i \in \{1, 2, \ldots, n\} \setminus \{k\} \). Then we have
\[
\{\ell'_1, \ell'_2, \ldots, \ell'_n\} \subseteq \{1, 2, \ldots, n + s - 1\}. \tag{23}
\]

Moreover, noticing that \( \{x_k, y_m, y_{\ell_k}\} \subseteq X_1 \cup X_2 \subseteq H^L \), then \( d_H(x_k, y_m) \leq L \) and \( d_H(x_k, y_{\ell_k}) \leq L \). So we can obtain
\[
d_H(x_k, y_m) - d_H(x_k, y_{\ell_k}) \leq L. \tag{24}
\]
And further we have
\[
d_S(X_1, X_2') = \sum_{i=1}^{n} d_H(x_i, y_{\ell_i}) + L(s - 1) \leq \sum_{i=1}^{n} d_H(x_i, y_{\ell_i}) + L(s - 1) = \sum_{i=1}^{n} d_H(x_i, y_{\ell_i}) - d_H(x_k, y_{\ell_k}) + d_H(x_k, y_m) + L(s - 1) \leq \sum_{i=1}^{n} d_H(x_i, y_{\ell_i}) + L + L(s - 1) = d_S(X_1, X_2),
\]
where the first inequality is obtained by (23) and (2), and the second inequality is obtained by (24).

Hence, we always have \( d_S(X_1, X_2') \leq d_S(X_1, X_2) \), which completes the proof.

**APPENDIX C**

**PROOF OF THEOREM 1**

By Definition 1 it is easy to see that for any two subsets \( X_1 \) and \( X_2 \) of \( H^L \), \( d_S(X_1, X_2) = d_S(X_2, X_1) \geq 0 \). Moreover, by Corollary 1 we can easily see that \( d_S(X_1, X_2) = 0 \) if and only if \( X_1 = X_2 \). So to prove that \( d_S(\cdot, \cdot) \) is a distance function, we only need to prove the triangle inequality, that is,
\[
d_S(X_1, X_2) \leq d_S(X_1, X_3) + d_S(X_2, X_3)
\]
for any three subsets \( X_1, X_2 \) and \( X_3 \) of \( H^L \). Without loss of generality, we can assume that \( |X_1| \leq |X_2| \). Then we have the following three cases.
Without loss of generality, we can assume
\[ d_s(X_1, X_2) = \sum_{i=1}^{n} d_H(x_i, z_i) + Ls, \]
\[ d_s(X_2, X_3) = \sum_{i=1}^{n} d_H(y_i, z_i) \]
and
\[ d_s(X_1, X_2) = \sum_{i=1}^{n} d_H(x_i, y_i) + Ls, \]
where \( s \geq 0 \) and \( \{\ell_1, \ell_2, \cdots, \ell_n\} \subseteq \{1, 2, \cdots, n + s\} \). Then we have
\[ d_s(X_1, X_2) = \sum_{i=1}^{n} d_H(x_i, y_i) + Ls \]
\[ \leq \sum_{i=1}^{n} d_H(x_i, y_i) + Ls \]
\[ \leq \sum_{i=1}^{n} (d_H(x_i, z_i) + d_H(y_i, z_i)) + Ls \]
\[ \leq \sum_{i=1}^{n} d_H(x_i, z_i) + Ls + \sum_{i=1}^{n+s} d_H(y_i, z_i) \]
\[ = d_s(X_1, X_3') + d_s(X_2, X_3') \]
\[ \leq d_s(X_1, X_3') + d_s(X_2, X_3'), \]
where the first inequality is obtained by (2) and the last inequality is obtained by Lemma 2.

Case 1. \( |X_1| \leq |X_2| \leq |X_3| \). In this case, we can fix a subset \( X_3' \subseteq X_3 \) of size \( |X_3'| = |X_2| \). Then by Lemma 2
\[ d_s(X_1, X_3') \leq d_s(X_1, X_3) \]
and
\[ d_s(X_2, X_3') \leq d_s(X_2, X_3). \]
So it suffices to prove that
\[ d_s(X_1, X_2) \leq d_s(X_1, X_3') + d_s(X_2, X_3'). \]

Without loss of generality, we can assume
\[ X_1 = \{x_1, \cdots, x_n\}, \]
\[ X_2 = \{y_1, \cdots, y_n, y_{n+1}, \cdots, y_{n+s}\}, \]
\[ X_3' = \{z_1, \cdots, z_n, z_{n+1}, \cdots, z_{n+s}\} \]
such that
\[ d_s(X_1, X_3') = \sum_{i=1}^{n} d_H(x_i, z_i) + Ls, \]
\[ d_s(X_2, X_3') = \sum_{i=1}^{n+s} d_H(y_i, z_i) \]
and
\[ d_s(X_1, X_2) = \sum_{i=1}^{n} d_H(x_i, y_i) + Ls, \]
where \( s \geq 0 \) and \( \{\ell_1, \ell_2, \cdots, \ell_n\} \subseteq \{1, 2, \cdots, n + s\} \). Then we have
\[ d_s(X_1, X_2) = \sum_{i=1}^{n} d_H(x_i, y_i) + Ls \]
\[ \leq \sum_{i=1}^{n} d_H(x_i, y_i) + Ls \]
\[ \leq \sum_{i=1}^{n} (d_H(x_i, z_i) + d_H(y_i, z_i)) + Ls \]
\[ \leq \sum_{i=1}^{n} d_H(x_i, z_i) + Ls + \sum_{i=1}^{n+s} d_H(y_i, z_i) \]
\[ = d_s(X_1, X_3') + d_s(X_2, X_3') \]
\[ \leq d_s(X_1, X_3') + d_s(X_2, X_3'), \]
where the first inequality is obtained by (2) and the last inequality is obtained by Lemma 2.

Case 2. \( |X_1| \leq |X_3| \leq |X_2| \). In this case, we can assume
\[ X_1 = \{x_1, \cdots, x_n\}, \]
\[ X_3 = \{y_1, \cdots, y_n, y_{n+1}, \cdots, y_{n+s}\}, \]
\[ X_2 = \{z_1, \cdots, z_n, z_{n+1}, \cdots, z_{n+s}, z_{n+s+1}, \cdots, z_{n+s+t}\} \]
such that
\[ d_s(X_1, X_3) = \sum_{i=1}^{n} d_H(x_i, y_i) + Ls, \]
\[ d_s(X_2, X_3) = \sum_{i=1}^{n+s} d_H(y_i, z_i) + Lt \]
and
\[ d_s(X_1, X_2) = \sum_{i=1}^{n} d_H(x_i, z_i) + L(s + t), \]
where \( s, t \geq 0 \) and \( \{ \ell_1, \ell_2, \cdots, \ell_n \} \subseteq \{ 1, 2, \cdots, n + s + t \} \). Then we have

\[
\begin{align*}
d_S(X_1, X_2) &= \sum_{i=1}^{n} d_H(x_i, z_{\ell_i}) + L(s + t) \\
&\leq \sum_{i=1}^{n} d_H(x_i, z_i) + L(s + t) \\
&\leq \sum_{i=1}^{n} (d_H(x_i, y_i) + d_H(y_i, z_i)) + L(s + t) \\
&\leq \sum_{i=1}^{n} d_H(x_i, y_i) + Ls + \sum_{i=1}^{n+s} d_H(y_i, z_i) + Lt \\
&= d_S(X_1, X_3) + d_S(X_2, X_3),
\end{align*}
\]

where the first inequality is obtained by (2).

**Case 3.** \( |X_3| \leq |X_1| \leq |X_2| \). In this case, we can assume

\[
X_3 = \{ x_1, \cdots, x_n \},
\]

\[
X_1 = \{ y_1, \cdots, y_n, y_{n+1}, \cdots, y_{n+s} \},
\]

\[
X_2 = \{ z_1, \cdots, z_n, z_{n+1}, \cdots, z_{n+s}, z_{n+s+1}, \cdots, z_{n+s+t} \}
\]

such that

\[
\begin{align*}
d_S(X_1, X_3) &= \sum_{i=1}^{n} d_H(x_i, y_i) + Ls, \\
d_S(X_2, X_3) &= \sum_{i=1}^{n} d_H(x_i, z_i) + L(s + t)
\end{align*}
\]

and

\[
\begin{align*}
d_S(X_1, X_2) &= \sum_{i=1}^{n+s} d_H(y_i, z_{\ell_i}) + Lt \\
&\leq \sum_{i=1}^{n+s} d_H(y_i, z_i) + Lt \\
&\leq \sum_{i=1}^{n+s} (d_H(x_i, y_i) + d_H(x_i, z_i)) + Lt \\
&\leq \sum_{i=1}^{n} d_H(x_i, y_i) + Ls + \sum_{i=1}^{n+s} d_H(y_i, z_i) + L(s + t) \\
&= d_S(X_1, X_3) + d_S(X_2, X_3),
\end{align*}
\]

where the first inequality is obtained by (2), and the third inequality is obtained from the simple fact that \( d_H(\cdot, \cdot) \leq L \).

For all cases, we have proved that

\[
\begin{align*}
d_S(X_1, X_2) &\leq d_S(X_1, X_3) + d_S(X_2, X_3),
\end{align*}
\]

So \( d_S(\cdot, \cdot) \) satisfies the triangle inequality.

By the above discussion, we proved that \( d_S(\cdot, \cdot) \) is a distance function over \( \mathcal{P}(\mathbb{L}^n) \).

**References**

[1] J. Davis, “Microvenus,” *Art J.*, 55, 70 (1996), doi:10.2307/777811
[2] G. M. Church, Y. Gao, and S. Kosuri, “Next-generation digital information storage in DNA,” *Science*, vol. 337, no. 6102, pp. 1628-1628, 2012.
[3] N. Goldman, P. Bertone, S. Chen, C. Dessissou, E. M. LeProust, B. Sipos, and E. Birney, “Towards practical, high-capacity, low-maintenance information storage in synthesized DNA,” *Nature*, vol. 494, no. 7435, pp. 77-80, 2013.
[4] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, “Robust chemical preservation of digital information on DNA in silica with error-correcting codes,” *Angew. Chem. Int. Ed.*, vol. 54, no. 8, pp. 2552-2555, 2015.
19

[5] M. Blawat, K. Gaedke-1, I. Hütter, X.-M. Chen, B. Turczyk, S. Inverso, B. W. Pruitt, G. M. Church, “Forward error correction for DNA data storage,” *Procedia Comput Sci*, vol. 80, pp. 1011-1022, 2016.

[6] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, “A Rewritable, Random-Access DNA-Based Storage System,” *Nature Scientific Reports*, 5(14138), 2015.

[7] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, “A DNA-based archival storage system,” in *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems, ACM*, pp. 637-649, 2016.

[8] Y. Erlich, and D. Zielinski, “DNA Fountain enables a robust and efficient storage architecture,” *Science*, vol. 355, no. 6328, pp. 950-954, 2017.

[9] W. Song, K. Cai, M. Zhang, and C. Yuen, “Codes with Run-Length and GC-Content Constraints for DNA-based Data Storage,” *IEEE Communications Letters*, 2018, DOI: 10.1109/LCOMM.2018.2866566

[10] K.A.S. Immink, and K. Cai, “Design of Capacity-Approaching Constrained Codes for DNA-Based Storage Systems,” *IEEE Communications Letters*, vol. 22, no. 2, pp. 224-227, 2018.

[11] A. Lenz, P. H. Siegel, A. W-Zeh, and E. Yaakobi, “Coding over Sets for DNA Storage,” 2018. Available: https://arxiv.org/abs/1801.04882

[12] J. Sima, N. Raviv, and J. Bruck, “On Coding over Sliced Information,” 2018. Available: https://arxiv.org/abs/1809.02716

[13] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, “Fundamental limits of DNA storage systems,” in *IEEE Int. Symp. Inform. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 3130-3134.

[14] 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.

[15] M. Langberg, M. Schwartz, and E. Yaakobi, “Coding for the $\ell_\infty$-Limited Permutation Channel,” *IEEE Trans. Inf. Theory*, vol. 63, no. 12, pp. 7676-7686, Dec. 2017.

[16] M. Kovačević, and V. Y. F. Tan, “Codes in the Space of Multisets — Coding for Permutation Channels with Impairments,” *IEEE Trans. Inf. Theory*, 2018, DOI: 10.1109/TIT.2017.2789292

[17] W. C. Huffman and V. Pless, *Fundamentals of Error-Correcting Codes*. Cambridge, U.K.: Cambridge Univ. Press, 2003

[18] S. Jukna, *Extremal Combinatorics*. New York: Springer-Verlag, 2001.