This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

On Algebraic Designing of DNA Codes with Biological and Combinatorial Constraints

Krishna Gopal Benerjee, *Member, IEEE,* and Adrish Banerjee, *Senior Member, IEEE*

***Abstract*—This paper presents constructions of DNA codes that satisfy biological and combinatorial constraints for DNA-based data storage systems. We introduce an algorithm that generates DNA blocks containing sequences that meet the required constraints for DNA codes.**

The constructed DNA sequences satisfy biological constraints: balanced GC-content, avoidance of secondary structures, and prevention of homopolymer runs. These sequences simultaneously satisfy combinatorial constraints that ensure differences among DNA sequences and their reverse and reverse-complement sequences. The DNA codes incorporate error correction through minimum Hamming distance requirements.

We establish a bijective mapping between algebraic structures and DNA sequences, providing construction of DNA codes with specified characteristics. Using this framework, we construct DNA codes based on error-correcting codes, including Simplex and Reed-Muller codes. These constructions ensure DNA sequences avoid secondary structures and homopolymer runs exceeding length three, which cause errors in DNA storage systems. Concatenated sequences maintain these properties.

The codes achieve non-vanishing code rates and minimum Hamming distances for large sequence lengths, demonstrating viability for DNA-based data storage systems.

***Index Terms*—DNA Data Storage, DNA Codes, GC-Content Constraint, Reversible Constraint, Reversible-Complement Constraint, Homopolymers, Secondary Structures.**

I. Introduction

DNA (DeoxyriboNucleic Acid) is carried by all known cellular life and some viruses. Biological information is stored in DNA for a long period, and DNA also helps transfer information between two generations. This motivates researchers to look at DNA for possible solutions to the long-term reliable storage of massive data. In 2012, Church *et al.* implemented a high-scale archival DNA-based data storage architecture on synthesized DNA [1]. Further, Goldman *et al.* proposed an encoding scheme for DNA-based data storage in 2013 [2]. DNA-based data storage is explored due to its incredible storage capacity, density, and durability [1], [2]. Details are discussed in survey papers [3]–[10].

A physical string of four *base pairs* or *nucleotides* Adenine (A), Thymine (T), Guanine (G), and Cytosine (C) is called DNA. In this paper, sequences made up of the four nucleotide bases A, C, G, and T are referred to as DNA strings. For any such DNA string, its complement is obtained by replacing each base with its complementary base, its reverse is formed by writing the string in reverse order, and its reverse-complement is the reverse of the complement string, where the Watson-Crick complements or simply complements, of DNA nucleotides are as follows: $A^c = T$, $G^c = C$, $T^c = A$, and $C^c = G$. For example, the DNA strings $TCCGAA$, $TTCGGA$, and $AAGCCT$ are the complement, reversible, and reversible-complement DNA strings of $AGGCTT$. In a wet lab, one can read and write physical DNA using DNA sequencing and a synthesizer [11]. Errors can occur when DNA is read and written in the lab. We list some of the properties of the DNA strings that help reduce errors while reading and writing a bunch of DNA strings as follows.

1. Avoiding Homopolymers: Errors such as nucleotide deletions, insertions, and substitutions occur frequently in DNA sequences. These errors are often attributed to the presence of homopolymers in DNA sequences [2], [12], [13]. Homopolymers are defined as consecutive repetitions of a nucleotide within a DNA sequence. A DNA sequence contains a *homopolymer of run-length* ℓ if a nucleotide appears at ℓ consecutive positions within the sequence. A DNA sequence is termed *ℓ-homopolymer-free* if it does not contain homopolymers with run-length greater than or equal to ℓ. For example, the DNA sequence $ACCCTGCAAAG$ contains homopolymers of run-length 3 and is 4-homopolymer-free.

2. Avoiding Secondary Structures: In DNA data storage, secondary structure constraints are relevant when sequences are single-stranded, such as during oligo synthesis and PCR. Primer-dimer formation is a key issue for primers [14]. Secondary structure is formed when nucleotides interact within a single-stranded DNA. A single-stranded DNA bends on itself and leads to some structures, called secondary structures [15]. Consecutive nucleotides that take part in a secondary structure of a single-stranded DNA are called the stem, and the number of successive nucleotides is known as stem length. Various loops and knots formed in secondary structures are discussed in [16]. Secondary structures, stem, and stem length for a DNA string is defined formally in Definition 1. If the stem length of each stem in the secondary structure of a DNA string is less than ℓ, then the DNA string is called *ℓ free-structures*. Examples of DNA strings with secondary structures are illustrated in Fig. 1. The secondary structure needs to be unfolded while reading the DNA. Therefore, additional energy and resources are needed to read the DNA, which increases costs. Milenkovic and Kashyap first constructed DNA codes that avoid secondary structures in their seminal paper [15]. The mfold Web Server [17] and Vienna [18] are generally used for RNA secondary structure prediction and analysis. It employs thermodynamics-based and dynamic programming algorithms to predict minimum free energy (MFE) structures, calculate base-pairing probabilities, visualize secondary structures, and design sequences for specified structures. In [19], [20], author

Authors are with the Department of Electrical Engineering, Indian Institute of Technology Kanpur, India email: {kgopal, adrish}@iitk.ac.in

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

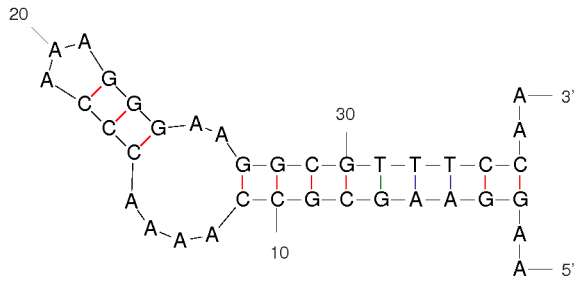

Fig. 1. For the DNA string $\mathbf{x}$ = $AAGGAAGCGCCAAAACCCAAAGGGAAGGCGTTTCCAA$ of length 37, the secondary structure with stems of length 3 and length 9 is predicted by The mfold Web Server [17] and Vienna [18]. Note that the sub-string $\mathbf{x}(16,18)$ = CCC attaches with the sub-string $\mathbf{x}(22,24)$ = GGG and forms a hairpin-like secondary structure with a stem of length three. Also, the sub-string $\mathbf{x}(3,11)$ = $GGAAGCGCC$ attaches with the sub-string $\mathbf{x}(27,35)$ = $GGCGTTTCC$ and forms a secondary structure with a stem of length nine.

has enumerating secondary structures that satisfy certain stability requirements using Context-Free Languages.

3. Reversible and Reversible-Complement Constraints: During the reading of a DNA string in the code, nonspecific hybridization can be avoided if this string differs sufficiently from (i) any other DNA string in the code (Hamming constraint), (ii) the reverse of any other distinct DNA string in the code (Reversible constraint), and (iii) the reverse-complement of any other distinct DNA string in the code (Reversible-Complement constraint) [11], [21]–[23]. Note that nonspecific hybridization refers to the unintended binding of DNA strands to non-target complementary sequences, resulting in erroneous hybridization events. Formal definitions of Reversible and Reversible-Complement constraints are given in Definition 3.

4. GC-Content Constraint: The *GC-content* or *GC-weight* of a DNA string is the total count of the nucleotide C and the nucleotide G in the DNA. For example, the GC-Content of the DNA string $AATACTCGTAGGTTTT$ is five. A balanced GC content in DNA codes is essential to maintain optimal thermal stability and prevent processing difficulties during sequencing and synthesis [13], [24]–[27]. Therefore, in this paper, we construct a set of DNA strings, each of length n, such that each DNA string has balanced GC-Content $\lfloor n/2 \rfloor$.

Therefore, DNA codes that avoid Homopolymers and secondary structures and satisfy Hamming, Reversible, Reversible-Complement, and GC-Content constraints are preferred.

Consequently, DNA codewords that circumvent homopolymer formations and secondary structure formations while adhering to Hamming distance requirements, reversibility constraints, reverse-complement restrictions, and GC-content balance are highly desirable for practical applications. Some work in this direction is as follows.

Related Literature: DNA codes with Reversible and Reversible-Complement constraints are constructed using BCH code on $GF(q^m)$ [22] and Reed-Muller code on $\mathbb{Z}_4[w]/\langle w^2-(2+2w)\rangle$ [21]. In addition, DNA codes with reverse-complement and GC-Content constraints are constructed using the field $GF(4)$ [28]–[30]. DNA codes in which DNA strings are free from consecutive repeats of blocks (a generalization of homopolymers) are studied in [31]. In addition, the DNA codes also satisfy Reversible, Reversible-Complement, and GC-Content constraints.

A significant body of research [32]–[40] has focused on error-correcting codes for DNA storage. These include codes for insertion/deletion errors [32], inversion errors [33], symbol errors [34], limited-magnitude probability errors [35], duplication errors [36], and coded trace reconstruction for efficient recovery from multiple noisy reads [37]. In [41], [42], authors studied codes under the Damerau metric that correct a single deletion and multiple adjacent transpositions, providing constructions, size bounds, and decoding algorithms. Burst-deletion/insertion correcting codes with improved redundancy and efficiency have also been proposed in [38]. Channel modeling and capacity analysis for DNA storage systems have been studied extensively in [43]–[46], including multi-draw channels corrupted by binary erasure [43], nanopore sequencing channels with intersymbol interference [46], and models addressing out-of-order sequencing in multi-molecule systems such as the shuffling [47], and shotgun sequencing [48] channels. In [49], authors investigated DNA sequence reconstruction from noisy substring observations, employing Markov type analysis to model the DNA storage channel and its relevance to synthesis and sequencing processes. Constrained DNA code design has been advanced through models capturing biochemical requirements such as balanced GC-content, secondary structure avoidance, and mutual uncorrelation. Kiah *et al.* [20], [50] pioneered asymmetric coding frameworks to mitigate synthesis and sequencing errors using combinatorial techniques. Yazdi *et al.* [14], [51] formulated weakly mutually uncorrelated codes that ensure balanced composition and large Hamming distances, addressing primer design challenges. Extensions to complex error models including insertions, deletions, and transpositions were provided by Gabrys *et al.* [42].

In addition to these foundational works, specialized coding techniques have been introduced to address specific challenges in DNA storage. Rank-Modulation codes for efficient data representation are explored in [52], while primer address codes for random access are constructed in [24]. Composite DNA coding methods for increasing alphabet size and correcting asymmetric errors are presented in [53], [54]. Guided scrambling and constrained codes for optimizing synthesis cycle counts in parallel strand synthesis are studied in [55]. Furthermore, capacity-achieving constrained codes balancing GC-Content and limiting homopolymer runs are proposed in [26], while enumerative coding methods for capacity-achieving GC-Content and prefix-constrained codes are developed in [56]. Also, improved constructions of Reed-Solomon (RS) and Generalized Reed-Solomon (GRS) codes meeting the half-Singleton bound are presented in [57]. In [58], authors show unconstrained coding (investigated for homopolymer or GC content restrictions) is more efficient against substitution errors than constrained coding. Apart from these two properties, we have considered several more properties, such as preventing secondary structures and holding reverse and reverse-complementing constraints simultaneously. Although constrained codes increase synthesis costs and lower density,

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

they ensure reliable recovery in high-error synthesis where unconstrained codes may fail.

These diverse contributions highlight the multifaceted nature of DNA storage coding research. They address fundamental biological constraints such as secondary structures, GC-balance, and homopolymer runs while advancing error correction techniques and channel modelling strategies to improve data reliability and storage efficiency.

Contributions: In this paper, DNA block sets are constructed from Algorithm 1 such that DNA strings defined over the DNA block set are ℓ free-structures. From Lemma 4, we have shown that any concatenation of those DNA strings is also ℓ free-structures. For DNA block set $\mathcal{A}_4 = \{AA, AC, CA, CC, CT, TC\}$ (see Remark 2), we have obtained the following results.

- As shown in Section III-A, any DNA string defined over the DNA block set $\mathcal{A}_4$ has the GC-Content half of the length (Lemma 5), and also, avoids the secondary structures (Proposition 2) and homopolymers of run-length of more than two (Lemma 7). Also, we have shown in Lemma 8 and Lemma 9 that any DNA code defined over the set $\mathcal{A}_4$ of DNA blocks holds the Reversible-Complement constraint. Apart from this, we have defined a bijective map $\psi : \mathbb{Z}_4^n \to \mathcal{A}_4$ and a distance d_ψ in Section IV. From the map property, we have given the condition on linear code $\mathscr{C}$ over $\mathbb{Z}_4$ in Lemma 10 so that the DNA code $\psi(\mathscr{C})$ holds the Reversible constraint. Further, from the distance-preserving property discussed in Lemma 11, we have given the parameters of obtained DNA codes in Theorem 1.
- Families of DNA codes such as Modified Simplex DNA Codes of Type 1 as given in Section VI-A, Modified Simplex DNA Codes of Type 2 as given in Section VI-B, and Reed-Muller Type DNA codes as discussed in Section VI-D are constructed that are 3 free-structures, 3 free-homopolymers, and also satisfy GC-Content and Reversible-Complement constraints along with Reversible constraint and each concatenated DNA string of these DNA codewords is again 3 free-structures. Also, all the constructed families of DNA codes have either a non-vanishing code rate or a non-vanishing relative minimum Hamming distance for large lengths.
- In particular, the relative minimum Hamming distance is approaching $1/2$ and $1/4$ for Modified Simplex DNA Codes of Type 1 and Modified Simplex DNA Codes of Type 2 for large lengths (see Remark 8 and Remark 10). And, for given $m - r$, the code rate approaches $1/2$ for Reed-Muller Type DNA codes for large lengths (see Remark 15).

Organisations: The basic preliminaries are discussed for DNA codes in Section II. Properties of DNA strings obtained from some DNA block sets are discussed in Section III. Further, DNA strings obtained from mapping between $\mathbb{Z}_4$ and a DNA block set are discussed in Section IV. Some examples of DNA codes are given in Section V, and then families of DNA codes are also obtained in Section VI. Finally, Section VII presents a discussion and comparison of different DNA codes.

II. Notations and Preliminary

In this section, we present some basic definitions, notations, and background results.

For any alphabet Σ of q symbols, any one-dimensional array, $(x_1\ x_2\ \dots\ x_n)$, of length n over Σ is known as a *string* $\mathbf{x}$, *i.e.*, $\mathbf{x} = (x_1\ x_2\ \dots\ x_n)$, where $x_i \in \Sigma$ for $i = 1, 2, \dots, n$. In particular, strings defined over the alphabet $\Sigma_{DNA} = \{A, T, C, G\}$ are called *DNA strings*. For any DNA string $\mathbf{y} = y_1 y_2 \dots y_n \in \Sigma_{DNA}^n$, strings $\mathbf{y}^c = y_1^c y_2^c \dots y_n^c$, $\mathbf{y}^r = y_n y_{n-1} \dots y_1$, and $\mathbf{y}^{rc} = y_n^c y_{n-1}^c \dots y_1^c$ are called *complement*, *reversible* and *reversible-complement* DNA strings, respectively. For simplicity, a DNA string of length n is represented by $\mathbf{x} = x_1 x_2 \dots x_n$, where $x_i \in \Sigma_{DNA}$ for $i = 1, 2, \dots, n$. For positive integers n and m, the *concatenation* of strings $\mathbf{x} = (x_1\ x_2 \dots\ x_n)$ and $\mathbf{y} = (y_1\ y_2 \dots\ y_m)$ is the string $\mathbf{xy} = (x_1\ x_2\ \dots\ x_n\ y_1\ y_2 \dots\ y_m)$ of length $n + m$. For a given string $\mathbf{x} = (x_1\ x_2\ \dots\ x_n)$ of length n and $1 \le i \le j \le n$, a string $(x_i\ x_{i+1}\ \dots\ x_j)$ is called *sub-string* of the string $\mathbf{x}$ and it is denoted by $\mathbf{x}(i, j)$. Similarly, for a DNA string $\mathbf{x} = x_1 x_2 \dots x_n$, a *DNA sub-string* is represented as $\mathbf{x}(i, j) = x_i x_{i+1} \dots x_j$. For any positive integers i, j, k, ℓ and any DNA string $\mathbf{x}$ of length n, two sub-strings $\mathbf{x}(i, j)$ and $\mathbf{x}(k, \ell)$ are known as *disjoint sub-strings* of $\mathbf{x}$ if $1 \le i \le j < k \le \ell \le n$. For example, the DNA sub-strings $\mathbf{x}(16, 18) = CCC$, $\mathbf{x}(22, 24) = GGG$, $\mathbf{x}(3, 11) = GGAAGCGCC$ and $\mathbf{x}(27, 35) = GGCGTTTCC$ of $\mathbf{x}$ as given in Fig. 1 are disjoint DNA sub-strings.

The Nussinov-Jacobson (NJ algorithm) folding algorithm can predict *approximately* secondary structure in a given DNA string [59]. Chemically active DNA strings become more stable when they release energy and form a secondary structure. Thus, one can predict secondary structure in a DNA string by computing a system property known as *free energy*. For a DNA string $\mathbf{x} = x_1 x_2 \dots x_n$, the free energy depends on the energy (known *interaction energy*) of pairing x_i and x_j $1 \le i < j \le n$. In the NJ algorithm, the interaction energy between the nucleotides x_i and x_j in the DNA string $\mathbf{x}$ is independent of any neighbor pairs. However, the interaction energy depends on the pairing of nucleotides x_i and x_j that contribute to the secondary structure. As given in [60], for any DNA string $\mathbf{x} = x_1 x_2 \dots x_n$, the frequently used values of interaction energy is

$$\beta(x_i, x_j) = \begin{cases} -5 & \text{for } (x_i, x_j) \in \{(C, G), (G, C)\}, \\ -4 & \text{for } (x_i, x_j) \in \{(T, A), (A, T)\}, \\ -1 & \text{for } (x_i, x_j) \in \{(T, G), (G, T)\}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For any sub-string $\mathbf{x}(i, j) = x_i x_{i+1} \dots x_j$ of the DNA string $\mathbf{x}$, the minimum free energy is given by $E_{i,j} = \min\{E_{i+1,j-1} + \beta(x_i, x_j), E_{i,k-1} + E_{k,j} : k = i + 1, \dots, j\}$, where $E_{\ell,\ell} = E_{\ell-1,\ell} = 0$ for $\ell = 1, 2, \dots, n$. The initial conditions follow the fact that any nucleotide of a DNA string does not bind with itself or immediate neighbors to form a secondary structure.

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

Following the mathematical definition of secondary structures given in [61], we define the secondary structure for a DNA string in Definition 1.

Definition 1 (**Secondary Structure**, [61])**.** *For any DNA string* $\mathbf{x} = x_1x_2\dots x_n \in \Sigma_{DNA}^n$ *of length* n*, a secondary structure is associated with a set* S *of pairs* (i,j) *for* $1 \le i < j \le n$ *such that both the nucleotides* x_i *and* x_j *are attached to each other. Any maximal subset of the set* S *associated with consecutive nucleotides in the secondary structures is known as a stem. Also, the number of those consecutive nucleotides is known as the stem length.*

Note that any set of pairs does not represent a valid secondary structure. Many options are rejected due to chemical and stereochemical constraints.

Example 1. *Consider a DNA string* $\mathbf{x} = x_1x_2\dots x_{37} = AAGGAAGCGCCAAAACCCAAAGGGAAGGCGTT-TCCAA$ *as given in Fig. 1. The secondary structure with two stems in* $\mathbf{x}$ *is associated with set* $S = \{(i,38-i), (j,40-j) : i = 3,4,\dots,11$ *and* $j = 16,17,18\}$*, where the two stems are associated with* $S_1 = \{(i,38-i) : i = 3,4,\dots, 11\}$ *and* $S_2 = \{(j,40-j) : j = 16,\ 17,18\}$ *of length* 9 *and* 3*, respectively. The secondary structure is also estimated by The mfold Web Server [17] and Vienna [18].*

Generally, a stem is defined as the maximal set of consecutive nested base pairs within a secondary structure. The classical stem definition covers reverse-complement binding structures but excludes complement binding structures due to biological nesting constraints. In contrast, our mathematical definition using maximal subsets of consecutive base pairs provides comprehensive coverage of both binding patterns. While the classical approach restricts analysis to biologically canonical stem formations, our formulation enables broader structural characterization including non-canonical parallel base-pairing configurations. This extended coverage facilitates more general secondary structure analysis in DNA code design applications.

Analysis of intersection energy shows that DNA sequences with a greater number of consecutive pairs in $\{(C,G),(T,A),(T,G),(G,C),(A,T),(G,T)\}$ are more susceptible to forming secondary structures with stems of any length, as our definition ensures every longer stem is composed of shorter nested stems as well. Therefore, we should reduce the number of these base pairs in the sequence to design DNA sequences that avoid secondary structures. To construct such DNA strings, Secondary-Complement (SC) DNA strings and Reversible-Secondary-Complement (RSC) DNA strings are formally defined in Definition 2 as follows.

Definition 2 (**SC and RSC DNA Strings**)**.** *For any DNA string* $\mathbf{x} \in \Sigma_{DNA}^{\ell}$*, the Secondary-Complement (SC) DNA string(s) and the Reversible-Secondary-Complement (RSC) DNA string(s) are* $\mathbf{x}^s = x_1^s x_2^s \dots x_\ell^s$ *and* $\mathbf{x}^{rs} = x_\ell^s x_{\ell-1}^s \dots x_1^s$*, respectively, where* $A^s = T$*,* $C^s = G$*,* $G^s = C$*,* $G^s = T$*,* $T^s = A$*, and* $T^s = G$*. Note, for any* $x \in \Sigma_{DNA}$*,* $\beta(x,x^s) < 0$ *and, for any* $y \in \Sigma_{DNA}$ *s.t.* $x \ne y^s$*,* $\beta(x,y) = 0$*.*

For example, consider a DNA string $\mathbf{x} = AATCGCC$. The SC DNA strings of $\mathbf{x}$ are $TTAGCGG$, $TTGGCGG$, $TTAGTGG$, $TTGGTGG$, and the RSC DNA strings of $\mathbf{x}$ are $GGCGATT$, $GGCGGTT$, $GGTGATT$, $GGTGGTT$. For $i,j = 1,2,\dots,n$ and $|j-i| > \ell$, two sub-strings $\mathbf{x}(i,i+\ell-1)$ and $\mathbf{x}(j,j+\ell-1)$ of a DNA string $\mathbf{x} = x_1x_2\dots x_n$ are called *ℓ-length disjoint Secondary-Complement and Reversible-Secondary-Complement DNA sub-strings* (*ℓ-length disjoint SC/RSC DNA sub-strings*) if $\mathbf{x}(i,i+\ell-1) \in \{\mathbf{x}(j,j+\ell-1)^s, \mathbf{x}(j,j+\ell-1)^{rs}\}$, where $\{\mathbf{z}^s, \mathbf{z}^{rs}\}$ is the set of all SC and RSC DNA strings of the DNA string $\mathbf{z}$. For example, in Fig. 1, sub-strings $\mathbf{x}(3,11)$ and $\mathbf{x}(27,35)$ are 9-length disjoint SC/RSC DNA sub-strings of the DNA string $\mathbf{x}$. Now, Lemma 1 and Lemma 2 are immediate.

Lemma 1. *For given two integers* ℓ *and* n *(*$1 \le \ell \le \lfloor n/2 \rfloor$*), consider a DNA string free from* ℓ*-length disjoint SC/RSC DNA sub-strings. Then, for any* $t \ge \ell$*, the DNA string does not contain* t*-length disjoint SC/RSC DNA sub-strings.*

Proof. For any DNA string of length n, if $t > \lfloor n/2 \rfloor$, then there do not exist two or more disjoint DNA sub-strings each of length t. Thus, the result holds for any integer $t > \lfloor n/2 \rfloor$. Now, for $t = \ell, \ell+1, \dots, \lfloor n/2 \rfloor$, consider a DNA string with two sub-strings $\mathbf{x} = x_1x_2\dots x_m$ and $\mathbf{y} = y_1y_2\dots y_m$ each of length m such that $\mathbf{x} \in \{\mathbf{y}^{rs}, \mathbf{y}^s\}$, where $x_j^{rs} = y_{m-j+1}$ $(j = 1,2,\dots,m)$ for $\mathbf{x} = \mathbf{y}^{rs}$, and $x_j^s = y_j$ $(j = 1,2,\dots,m)$ for $\mathbf{x} = \mathbf{y}^s$.

- Case 1 ($\mathbf{x} = \mathbf{y}^{rs}$): Now, for any positive integers ℓ $(\le m)$ and i $(\le m-\ell)$, if possible then consider two sub-strings $\mathbf{x}_1 = x_ix_{i+1}\dots x_{i+\ell-1}$ and $\mathbf{y}_1 = y_{m-i+1}y_{m-i+2}\dots y_{m-i-\ell+2}$ of $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x}_1^s \ne \mathbf{y}_1^{rs}$. But it is not possible because $x_j^{rs} = y_{m-j+1}$. Thus, from contradiction, the result is proved for the case $\mathbf{x} = \mathbf{y}^{rs}$.
- Case 2 ($\mathbf{x} = \mathbf{y}^s$): One can prove the result for this case using similar arguments as given in Case 1.

Thus, we obtain the result from both cases. □

Lemma 2. *An* ℓ *free-homopolymers DNA string does not contain homopolymers of run-lengths greater than* $\ell - 1$*.*

Proof. A DNA string $\mathbf{x} = x_1x_2\dots x_n \in \Sigma_{DNA}^n$ is free from homopolymers of run-length ℓ if and only if, for all $j \in \{1,2,\dots,n-\ell+1\}$, there exists some $i \in \{j, j+1,\dots,j+\ell-1\}$ such that $x_i \ne x_j$. For any positive integer $m \ge \ell$, $\{j,j+1,\dots,j+\ell-1\} \subseteq \{j,j+1,\dots,j+m-1\}$ and $\{1,2,\dots,n-m+1\} \subseteq \{1,2,\dots,n-\ell+1\}$. Therefore, for all $j \in \{1,2,\dots,n-m+1\}$, there exists some $i \in \{j,j+1,\dots,j+m-1\}$ such that $x_i \ne x_j$. Thus, the DNA string $\mathbf{x}$ is also ℓ free-homopolymers. But, m is an arbitrary positive integer more than equal to ℓ. □

For any positive integers n, M and d, an (n,M,d) *code* is a subset $\mathscr{C} \subset \Sigma^n$ of size M with the minimum distance $d = \min\{d(\mathbf{a},\mathbf{b}) : \mathbf{a} \ne \mathbf{b}$ and $\mathbf{a},\mathbf{b} \in \mathscr{C}\}$, where $d(\mathbf{a},\mathbf{b})$ is the distance between the codewords $\mathbf{a}$ and $\mathbf{b}$ in $\mathscr{C}$. In the case of Hamming, the distance between $\mathbf{a}$ and $\mathbf{b}$ and the minimum Hamming distance (MHD) are denoted by $d_H(\mathbf{a},\mathbf{b})$ and d_H, respectively. Consider a code $\mathscr{C}$ defined over Σ_q with q symbols if, for any $\mathbf{x},\mathbf{y} \in \mathscr{C}$ and $a,b \in \Sigma_q$, $(a\cdot\mathbf{x})+(b\cdot\mathbf{y}) \in \mathscr{C}$ then

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

the code $\mathscr{C}$ is known as a linear code, where Σ_q is either a ring or a field of size q with additive and multiplicative operations '$+$' and '$\cdot$', respectively. For any matrix G over Σ_q, $\langle G\rangle = \{(a\cdot\mathbf{x})+(b\cdot\mathbf{y}) : \mathbf{x} \text{ and } \mathbf{y} \text{ are two rows of } G \text{ and } a,b\in\Sigma_q\}$. In particular, a matrix G defined over the ring $\mathbb{Z}_4$ is called type $\{k_0,k_1\}$ if G can be deduced into $\begin{pmatrix} I_{k_0} & B_{1,2} & B_{1,3} \\ \mathbf{0}_{2,1} & 2I_{k_1} & 2B_{2,3}\end{pmatrix}$, where I_m are identity matrices of order m for $m\in\{k_0,k_1\}$, $\mathbf{0}_{2,1}$ is a zero matrix with k_0 rows and k_1 columns, and $B_{i,j}$ are matrices with k_{j-1} columns and k_{i-1} rows over $\mathbb{Z}_4$ for $1\le i<j\le 3$. For any $\{k_0,k_1\}$ type matrix G over $\mathbb{Z}_4$, the size of the set $\langle G\rangle$ is $4^{k_0}2^{k_1}$ [62]. An (n,M,d_H) *DNA code* $\mathscr{C}_{DNA}$ is a code with parameter (n,M,d_H) over the DNA alphabet $\Sigma_{DNA}=\{A,T,C,G\}$. Now, for any (n,M,d_H) DNA code, the *code rate* is given by $\frac{1}{n}\log_4 M$, and the *relative minimum Hamming distance* (RMHD) is $\frac{d_H}{n}$.

For DNA codes, constraints given in Definition 3 are crucial for reducing sequencing errors and enhancing data integrity. Definition 3 presents the key constraints commonly considered in the design of DNA codes.

Definition 3 (**Constraints of DNA Codes**). *An (n,M,d_H) DNA code $\mathscr{C}_{DNA}$*

1) *holds Reversible constraint (R constraint) if $d_H^r\ge d_H$,*
2) *holds Reversible-Complement constraint (RC constraint) if $d_H^{rc}\ge d_H$,*
3) *holds GC-Content constraint if GC-Content of all DNA codewords are $\lfloor n/2\rfloor$,*
4) *called ℓ free-structures if all DNA codewords are ℓ free-structures, and*
5) *called ℓ free-homopolymers if all DNA codewords are ℓ free-homopolymers,*

where $d_H^r=\min\{d_H(\mathbf{x},\mathbf{y}^r):\mathbf{x},\mathbf{y}\in\mathscr{C}_{DNA}\ s.t.\ \mathbf{x}\neq\mathbf{y}^r\}$ and $d_H^{rc}=\min\{d_H(\mathbf{x},\mathbf{y}^{rc}):\mathbf{x},\mathbf{y}\in\mathscr{C}_{DNA}\ s.t.\ \mathbf{x}\neq\mathbf{y}^{rc}\}$.

For example, consider an $(5,2,5)$ DNA code $\mathscr{C}_{DNA}=\{ACACA,GAGAG\}$ that holds R, RC, and GC-Content constraint, and also, the DNA code is 1 free-structures and 1 free-Homopolymers. Any DNA code $\mathscr{C}_{DNA}$ contains reversible DNA strings, *i.e.*, for each $\mathbf{x}\in\mathscr{C}_{DNA}$, $\mathbf{x}^r\in\mathscr{C}_{DNA}$ then the DNA code holds R constraint. And therefore, for any DNA code $\mathscr{C}_{DNA}$, the DNA code $\mathscr{C}_{DNA}\cup\mathscr{C}^r_{DNA}$ holds R constraint, where $\mathscr{C}^r_{DNA}=\{\mathbf{x}^r:\mathbf{x}\in\mathscr{C}_{DNA}\}$. Similarly, any DNA code $\mathscr{C}_{DNA}$ contains complement and reversible DNA strings, *i.e.*, for each $\mathbf{x}\in\mathscr{C}_{DNA}$, DNA strings $\mathbf{x}^r,\mathbf{x}^c\in\mathscr{C}_{DNA}$ then the DNA code holds R and RC constraints. And thus, for any DNA code $\mathscr{C}_{DNA}$, the DNA code $\mathscr{C}_{DNA}\cup\mathscr{C}^r_{DNA}\cup\mathscr{C}^c_{DNA}\cup\mathscr{C}^{rc}_{DNA}$ holds R and RC constraints, where $\mathscr{C}^c_{DNA}=\{\mathbf{x}^c:\mathbf{x}\in\mathscr{C}_{DNA}\}$ and $\mathscr{C}^{rc}_{DNA}=\{\mathbf{x}^{rc}:\mathbf{x}\in\mathscr{C}_{DNA}\}$.

III. DNA Block sets and their Properties

For any given integer ℓ, we define the DNA block set as a subset of Σ^ℓ_{DNA}. We discuss the properties of DNA block sets $\mathcal{A}$ so that DNA strings defined over the DNA block set ensure the required properties. In this paper, note that elements of $\mathcal{A}$ are referred to solely as blocks, the DNA codewords derived from $\mathcal{A}$ are not termed blocks.

Lemma 3 describes a mathematical property of DNA strings for secondary structures with a given stem length. Using this property, we can design DNA sequences that do not form such structures. The Lemma 3 is as follows.

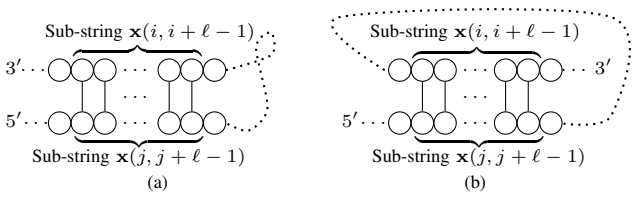

Fig. 2. For any DNA string $\mathbf{x}$ with secondary structure, possible binding types are shown. For sub-strings $\mathbf{x}(i,i+\ell-1)$ and $\mathbf{x}(j,j+\ell-1)$ of the DNA string $\mathbf{x}$, if binding occurs as shown in (a) then $\mathbf{x}(i,i+\ell-1)=\mathbf{x}(j,j+\ell-1)^{rs}$, and if binding occurs as given in (b) then $\mathbf{x}(i,i+\ell-1)=\mathbf{x}(j,j+\ell-1)^{s}$. DNA Secondary Structure Binding Types: (a) Reverse-complement binding $\mathbf{x}(i,i+\ell-1)=\mathbf{x}(j,j+\ell-1)^{rs}$; (b) Complement binding $\mathbf{x}(i,i+\ell-1)=\mathbf{x}(j,j+\ell-1)^{s}$.

Lemma 3. *For given integers n and ℓ ($1\le\ell\le\lfloor n/2\rfloor$), if a DNA string of length n forms a secondary structure that consists of the stem(s) of length ℓ then the DNA string contains at-least two disjoint sub-strings each of length ℓ such that both sub-strings are either Secondary-Complement or Reversible-Secondary-Complement of each other.*

Proof. Consider a DNA string $\mathbf{x}=x_1x_2\dots x_n$ that forms a secondary structure associated with $S=\{(i,f_s(i)):i\in I\}$ for some $I\subset\{1,2,\dots,n\}$. The secondary structure consists of m stems, and thus, one can find a partition $\{I_t:t=1,2,\dots,m\}$ of the set I. For each $t=1,2,\dots,m$, the set $S_t=\{(i,f_s(i)):i\in I_t\}\subset S$ is associated with a stem of length ℓ_t, where $I=\bigcup_{t=1}^m I_t$ and $|S_t|=|I_t|=\ell_t$. For the given I_t, if $j=\min\{i:i\in I_t\}$ and $k=\min\{f_s(i)):i\in I_t\}$ then there are two sub-strings $\mathbf{x}(j,j+\ell_t-1)=x_jx_{j+1}\dots x_{j+\ell_t-1}$ and $\mathbf{x}(k,k+\ell_t-1)=x_kx_{k+1}\dots x_{k+\ell_t-1}$ such that both sub-strings attach pairwise to each other and it forms the stem of length ℓ_t. Note that, any nucleotide does not bind with itself or its immediate neighbours, and therefore, $|k-j|>\ell_t$. Clearly, from Definition 2, $x_i=x^s_{f_s(i)}$ for $i=j,j+1,\dots,j+\ell_t-1$. Now, as shown in Fig. 2, there are two cases as follows.

- **Case 1**: As shown in Fig. 2(a), if sub-string $\mathbf{x}(j,j+\ell_t-1)$ attaches pairwise with the sub-string $\mathbf{x}(k,k+\ell_t-1)$ in reversible direction then $(i,f_s(i))\in\{(j,k+\ell_t-1),(j+1,k+\ell_t-2),\dots,(j+\ell_t-1,k)\}$, and thus, $f_s(i)=j+k+\ell_t-1-i$. In this case, from Definition 2, $\mathbf{x}(j,j+\ell_t-1)=\mathbf{x}(k,k+\ell_t-1)^{rs}$.
- **Case 2**: As given in Fig. 2(b), if sub-string $\mathbf{x}(j,j+\ell_t-1)$ attaches pairwise with the sub-string $\mathbf{x}(k,k+\ell_t-1)$ in same direction then $(i,f_s(i))\in\{(j,k),(j+1,k+1),\dots,(j+\ell_t-1,k+\ell_t-1)\}$ and $f_s(i)=k-j+i$. In this case, from Definition 2, $\mathbf{x}(j,j+\ell_t-1)=\mathbf{x}(k,k+\ell_t-1)^{s}$.

So, from both cases and Definition 2, it follows the result. □

From Lemma 1 and Lemma 3, one can directly conclude the Proposition 1.

Proposition 1. *For given positive integers n and ℓ ($\le\lfloor n/2\rfloor$), consider a DNA string of length n such that the string is $\ell+1$ free-structures. Then any secondary structure with stems of length more than ℓ does not exist in the DNA string.*

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

Now, one can immediately obtain a set over Σ_{DNA} such that any string defined over the set is free from Reversible-Secondary-Complement sub-strings and also free from Secondary-Complement sub-strings of a certain length, and therefore, from Proposition 1, it avoids the secondary structures. Now, one can obtain conditions on such DNA block set in Lemma 4 as follows.

Lemma 4. *For any integers t, n and ℓ ($2 \leq \ell \leq \lfloor t/2 \rfloor \leq n$), and a set $\mathscr{A} \subset \Sigma_{DNA}^t$, if all DNA strings in $\mathscr{A}^4$ do not contain ℓ-length disjoint SC/RSC DNA sub-strings then all DNA strings in $\mathscr{A}^n$ are ℓ free-structures, where $\mathscr{A}^k$ denotes the set of all possible concatenations of k strings from $\mathscr{A}$.*

Proof. For $i = 1, 2, 3, 4$, consider $\mathbf{x}_i = x_{i,1}x_{i,2}\dots x_{i,t}$ in $\mathscr{A}$, and thus the length of the DNA string $\mathbf{z} = \mathbf{x}_1\mathbf{x}_2\mathbf{x}_3\mathbf{x}_4 \in \mathscr{A}^4$ is $4t$, where the DNA string is free from ℓ-length disjoint SC/RSC DNA sub-strings. The condition $2 \leq \ell \leq \lfloor t/2 \rfloor$ ensures the existence of two ℓ-length and disjoint DNA sub-strings of $\mathbf{x}_i$. Now, for any $1 \leq j \leq 3$, $1 \leq r \leq t - \ell$ and $k \in \{1, 2, \dots, t\}$, consider ℓ-length sub-strings $\mathbf{y}_{j,k} = x_{j,t-k}x_{j,t-k+1}\dots x_{j,t}x_{j+1,1}x_{j+1,2}\dots x_{j+1,\ell-k-1}$ and $\mathbf{z}_{i,r} = x_{i,r}x_{i,r+1}\dots x_{i,r+\ell-1}$ of the DNA string $\mathbf{z} \in \mathscr{A}^4$. Also, consider a set P_ℓ of all these sub-strings. The DNA string $\mathbf{z}$ does not contain ℓ-length disjoint SC/RSC DNA sub-strings. It ensures the following properties.

1) DNA string $\mathbf{x}_i \in \mathscr{A}$ is free from ℓ-length disjoint SC/RSC DNA sub-strings for each i.
2) For any $1 \leq i_1, i_2 \leq 4$ and $1 \leq r_1, r_2 \leq t - \ell$, DNA sub-strings $\mathbf{z}_{i_1,r_1} \neq \mathbf{z}^s_{i_2,r_2}$ and $\mathbf{z}_{i_1,r_1} \neq \mathbf{z}^{rs}_{i_2,r_2}$.
3) For any $1 \leq j_1, j_2 \leq 3$ and $k_1, k_2 \in \{1, 2, \dots, t\}$, DNA sub-strings $\mathbf{y}_{j_1,k_1} \neq \mathbf{y}^{rs}_{j_2,k_2}$ and $\mathbf{y}_{j_1,k_1} \neq \mathbf{y}^{s}_{j_2,k_2}$.
4) For any $1 \leq i \leq 4$, $1 \leq r \leq t - \ell$, $1 \leq j \leq 3$ and $k \in \{1, 2, \dots, t\}$, DNA sub-strings $\mathbf{z}_{i,r} \neq \mathbf{y}^{s}_{j,k}$ and $\mathbf{z}_{i,r} \neq \mathbf{y}^{rs}_{j,k}$.

Since, for $i = 1, 2, 3, 4$, chosen DNA strings $\mathbf{x}_i$ are arbitrary, so properties given in 1, 2, 3 and 4 hold for any combination of DNA strings from $\mathscr{A}$. Now, for any DNA string in $\mathscr{A}^n$, any sub-string of length ℓ is also in the set P_ℓ, and therefore, from properties 1, 2, 3 and 4 as discussed above in this proof, are neither Secondary-Complement nor Reversible-Secondary-Complement of each other. Thus, the result holds from Lemma 1 and Lemma 3. □

Note that, for a given integer ℓ, Lemma 4 ensures the existence of a set of DNA blocks such that DNA strings and concatenation of those strings defined over the set are ℓ free-structures.

Thus, for given ℓ and t, such sets can be found using a computational approach. One possible computational approach is given in Algorithm 1.

Remark 1. *Algorithm 1 compares pairs of disjoint ℓ-length substrings within DNA sequences of length $4t$, resulting in approximately $\mathcal{O}((4t-2\ell)^2)$ comparisons per sequence. Considering all DNA strings containing at least one fixed substring in T^4, the number of such strings grows combinatorially with $|T|$, dominated by $\mathcal{O}((|T|-1)^3)$, where $|T|$ represents the size of the set T. Given the upper bound $|T| \leq 6 \cdot 4^{t-2}$, the overall complexity of substring comparisons in the algorithm is $\mathcal{O}((4t-2\ell)^2\, 4^{3t-6})$. For more detail, refer Appendix A.*

Algorithm 1 DNA block set $\mathscr{A}$ construction algorithm

Input: Parameters ℓ and t ($2 \leq \ell \leq \lfloor t/2 \rfloor$)
Output: DNA block set $\mathscr{A} \subseteq \Sigma_{DNA}^t$ such that all DNA string defined over $\mathscr{A}$ are ℓ free-structures
Initialisation : Set $S = \Sigma_{DNA}^t$, set $\mathscr{A} = \emptyset$ and $z \in S$
while $S \neq \emptyset$ **do**
  Compute $T = \mathscr{A} \cup \{z\}$ and $S = S \backslash \{z\}$ for $z \in S$
  if (All strings that contain at least one occurrence of substring z in T^4 are free from ℓ-length disjoint SC/RSC DNA substrings.) **then**
    Update $\mathscr{A} = \mathscr{A} \cup \{z\}$
  end if
end while
return $\mathscr{A}$

For $t = 2, 3$ and $\ell = 3$, the set $\mathscr{A}$ is calculated using Algorithm 1, and those DNA block sets are listed in Table I. Any DNA string defined over the DNA block set $\mathscr{A}$ as listed in Table I, has the following properties.

- For any DNA string defined over the DNA block set $\mathscr{A}$, the sum of As and Cs is lot more than the sum of Gs and Ts. Therefore, there are very few pairs $(x, x^s) \in \Sigma_{DNA}^2$ such that $\beta(x, x^s) < 0$. Thus, the property helps to avoid secondary structures in these DNA strings.
- At least one immediate neighbour of each nucleotide T is C and of each G is A, and it controls the stem length of secondary structures if it exists in the DNA string.

Further, avoiding homopolymers is not ensured by DNA strings over the DNA block sets obtained from Algorithm 1 (see DNA block set $\mathscr{A}$ in Table I). Therefore, we obtained sub-sets of those DNA block sets so that DNA strings defined over the sub-sets avoid homopolymers and secondary structures. The following Remark 2 explores the properties of specific subsets of DNA strings, focusing on their ability to avoid secondary structures.

Remark 2. *For $t = 2$ and $\ell = 3$, consider the set $\mathcal{A} = \{AA, AC, CA, CC, CT, TC\} \subset \Sigma_{DNA}^2$ and the sub-set $\mathcal{A}_4 = \{AC, CA, TC, CT\} \subset \mathcal{A}$. All DNA strings in $\mathcal{A}^4$ are 3 free-structures, and therefore, from Lemma 4, each DNA string in $\mathcal{A}^n$ is 3 free-structures. Similarly, from Lemma 4, all DNA strings in $\mathcal{A}_4^n$ are also 3 free-structures.*

Now, for the remaining part of the paper, we fix the notations $\mathcal{A} = \{AA, AC, CA, CC, CT, TC\}$, and $\mathcal{A}_4 = \{AC, CA, TC, CT\}$. Note that $\mathcal{A}_4 \subset \mathcal{A}$. Note that $\mathcal{A}_4$ is a set with four elements, where each element is a DNA string of length two.

A. DNA sub-set $\mathcal{A}_4$ and properties

For any DNA strings defined over $\mathcal{A}_4$, properties such as avoiding secondary structures and homopolymers, GC-Content, and Hamming distance are discussed in this section.

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

TABLE I
DISTINCT DNA BLOCK SETS ARE LISTED, WHERE ANY DNA STRING DEFINED OVER THE SET $\mathscr{A}$ IS 3 FREE-STRUCTURES. SUBSETS $\mathcal{A}$ OF THOSE ALPHABETS $\mathscr{A}$ ARE ALSO LISTED, WHERE DNA STRINGS DEFINED OVER THE SUBSET $\mathcal{A}$ ARE 3 FREE-STRUCTURES AND ALSO AVOID HOMOPOLYMERS.

t	DNA block set $\mathscr{A}$	$\frac{1}{t}\log_4(\lvert\mathscr{A}\rvert)$	Set $\mathcal{A}$ ($\subseteq \mathscr{A}$)	$\frac{1}{t}\log_4(\lvert\mathcal{A}\rvert)$
2	$\{AA, AC, CA, CC, TC, CT\}$	0.6462	$\{AC, CA, TC, CT\}$	0.5000
3	$\{AAA, AAC, ACA, CAA, ACC, CAC,$ $CCA, CCC, CTA, CTC, ATC, AGC, \quad AGA, CGA\}$	0.6346	$\{ACA, CAC, CTA,$ $CTC, ATC, AGC,$ $AGA, CGA\}$	0.5000
	$\{AAA, AAC, ACA, CAA, ACC, CAC,$ $CCA, CCC, ATC, ACT, CTC, CCT,$ $TCA, CTA, TCC\}$	0.6511	$\{ACA, CAC, ATC,$ $ACT, CTC, CCT,$ $TCA, CTA\}$	0.5000
4	$\{AAAA, AAAC, AACA, ACAA, CAAA,$ $AACC, ACAC, ACCA, CAAC, CACA,$ $CCAA, CCCA, CCAC, CACC, ACCC,$ $CCCC, CTAA, CTAC, CTCA, CTCC,$ $CCTA, ACTC, ACTA, ATCA, ATCC,$ $AATC, CATC, AGAA, AGAC, AGCA,$ $AGCC, AAGA, AAGC, CAGA, CAGC,$ $ACGA, CCGA, CGAA, CGAC\}$	0.6607	$\{ACAC, ACCA,$ $CAAC, CACA,$ $CTAC, CTCA,$ $ACTC, ACTA,$ $ATCA, CATC,$ $AGAC, AGCA,$ $CAGA, CAGC,$ $ACGA, CGAC\}$	0.5000
	$\{AAAA, AAAC, AACA, ACAA, CAAA,$ $AACC, ACAC, ACCA, CAAC, CACA,$ $CCAA, CCCA, CCAC, CACC, ACCC,$ $CCCC, CCCT, ACCT, CACT, AACT,$ $CTAC, CCTC, CCTA, ACTC, ACTA,$ $CTAA, CTCA, CTCC, TCAA, TCAC,$ $TCCA, TCCC, ATCC, ATCA, AATC,$ $CATC, CTTC, TCCT\}$	0.6560	$\{ACAC, ACCA,$ $CAAC, CACA,$ $ACCT, CACT,$ $ACTC, ACTA,$ $CTAC, CTCA,$ $TCAC, TCCA,$ $ATCA, CATC,$ $CTTC, TCCT\}$	0.5000

1) *Secondary structures property*: From Remark 2, the property of avoiding secondary structures in DNA strings defined over $\mathcal{A}_4$ is discussed in Proposition 2 as follows.

Proposition 2. *Any DNA string defined over $\mathcal{A}_4$ is 3 free-structures.*

2) *GC-Content property*: For any DNA string over $\mathcal{A}_4$, the GC-Content property is given in following lemma.

Lemma 5. *The GC-Content of any DNA string defined over $\mathcal{A}_4$ is half of its length.*

Proof. For any $x \in \mathcal{A}_4$, the element x is associated with two nucleotides, and the GC-Content of x is one. Thus, any DNA string in $\mathcal{A}_4^n$ has the length $2n$ and GC-Content n. □

Now, conditions on DNA codes with GC-Content constraint are given in Lemma 6.

Lemma 6. *Any DNA code defined over $\mathcal{A}_4$ holds GC-Content constraint.*

Proof. The result holds from Lemma 5, and the definition of GC-Content constraint. □

3) *Homopolymer property*: For any DNA strings over $\mathcal{A}_4$, the property of avoiding homopolymers is given in the following lemma and remark.

Lemma 7. *Each codeword of any DNA code defined over $\mathcal{A}_4 \subset \mathcal{A}$ is 3 free-homopolymers.*

Proof. For $y_1, y_2, y_3, y_4 \in \Sigma_{DNA}$, if $y_1y_2, y_3y_4 \in \mathcal{A}_4$ then $y_1 \neq y_2$ and $y_3 \neq y_4$. For any DNA string $\mathbf{x} = x_1x_2 \dots x_{2n} \in \mathcal{A}_4^n$, observe that $x_{2i-1} \neq x_{2i}$ for each $i = 1, 2, \dots, n$. Therefore, any three or more consecutive nucleotides are not the same. Thus, the DNA string $\mathbf{x}$ defined over $\mathcal{A}_4$ is 3 free-homopolymers. □

Remark 3. *Any DNA string obtained by concatenating DNA strings defined over $\mathcal{A}_4$ is also defined over $\mathcal{A}_4$. Therefore, from Proposition 2, and Lemma 7, any such concatenated DNA strings are 3 free-structures and 3 free-homopolymers. Again, from Lemma 5, the GC-Content of those concatenated DNA strings are also half of their respective lengths.*

4) *Reversible-Complement property*: For any DNA code defined over $\mathcal{A}_4$, the reversible-complement property is ensured in Lemma 8 and Lemma 9.

Lemma 8. *Any (n, M, d_H) DNA code over the DNA block set $\mathcal{A}_4$ holds the RC constraint, if $d_H \leq n$.*

Proof. Consider a DNA code $\mathscr{C}_{DNA}$ over the DNA block set $\mathcal{A}_4$. For any $x, y \in \mathcal{A}_4$, the distance $d_H(x, y^{rc}) \geq 1$. Thus, for any DNA codewords $\mathbf{x} = x_1x_2 \dots x_n$ and $\mathbf{y} = y_1y_2 \dots y_n$ in $\mathscr{C}_{DNA}$, the reversible-complement DNA string $\mathbf{y}^{rc} = y_n^{rc}y_{n-1}^{rc} \dots y_1^{rc}$. Therefore, the distance $d_H(\mathbf{x}, \mathbf{y}^{rc}) = \sum_{i=1}^{n} d_H(x_i, y_{n-i+1}^{rc}) \geq n \geq d_H$, and it holds the result. □

Now, a condition on DNA codes with RC constraint is given in Lemma 9.

Lemma 9. *For any $(2n, M, d_H)$ DNA code $\mathscr{C}_{DNA}$ over $\mathcal{A}_4$, $\mathscr{C}_{DNA} \cup \mathscr{C}_{DNA}^c$ is an $(2n, 2M, d_H^*)$ DNA code, where $d_H^* = \min\{d_H, d_H^c\}$, and $d_H^c = \min\{d_H(\mathbf{x}, \mathbf{y}^c) : \mathbf{x}, \mathbf{y} \in \mathscr{C}_{DNA}\}$. If $d_H \leq n$ then $d_H^* = d_H$.*

Proof. The result holds from Lemma 8 and facts that (i) for any $\mathbf{x} \in \mathscr{C}_{DNA}$, the length of $\mathbf{x}$ and $\mathbf{x}^c$ are same, (ii) for any DNA code $\mathscr{C}_{DNA}$ over $\mathcal{A}_4$, $\mathscr{C}_{DNA} \cap \mathscr{C}_{DNA}^c = \emptyset$, (iii) $d_H(\mathbf{x}, \mathbf{y}) = d_H(\mathbf{x}^c, \mathbf{y}^c)$ for any $\mathbf{x}, \mathbf{y} \in \Sigma_{DNA}^n$, and (iv) $d_H(x, y) \geq 1$ for any $x, y \in \mathcal{A}_4$. □

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

IV. DNA CODES OVER $\mathbb{Z}_4$

In this section, we defined the map ψ and analyzed the properties of the resulting DNA sequences and corresponding DNA codes. This provides a foundation for subsequent analysis of DNA codes with these all properties.

For any positive integer n and the set $\mathcal{A}_4 = \{AC, CA, TC, CT\} \subset \mathcal{A}$, define a bijective map $\psi : \mathbb{Z}_4^n \to \mathcal{A}_4$, *s.t.*, $\psi(0) = CT$, $\psi(1) = CA$, $\psi(2) = TC$, $\psi(3) = AC$. For any $\mathbf{x} = (x_1\ x_2\ \dots\ x_n) \in \mathbb{Z}_4^n$, consider $\psi(\mathbf{x}) = \psi(x_1)\psi(x_2)\dots\psi(x_n)$. For example, quaternary string $(0\ 1\ 1\ 3\ 2)$ is encoded into $CTCACAACTC$. For each $\mathscr{C} \subset \mathbb{Z}_4^n$, $\psi(\mathscr{C}) = \{\psi(\mathbf{x}) : \mathbf{x} \in \mathscr{C}\}$. Now, the properties of the map ψ are listed as follows.

Property 1. *Any n length string $\mathbf{x}$ over $\mathbb{Z}_4$ is encoded into a $2n$ length DNA string since $\mathcal{A}_4^n \subset \Sigma_{DNA}^{2n}$. As an example, the length of the DNA string $\psi((0\ 0\ 1)) = CTCTCA$ is six.*

Property 2. *For each $x \in \mathbb{Z}_4$, $\psi(x)^r = \psi(2 + x)$. Thus, for any $\mathbf{x} = (x_1\ x_2\ \dots\ x_n) \in \mathbb{Z}_4^n$, $\psi(\mathbf{x})^r = (\psi(x_1)\psi(x_2)\dots\psi(x_n))^r = \psi(x_n)^r\psi(x_{n-1})^r\dots\psi(x_1)^r = \psi(2 + x_n)\psi(2 + x_{n-1})\dots\psi(2 + x_1)$. For example, consider $\psi((0\ 0\ 1)) = CTCTCA$. Than, $\psi((0\ 0\ 1))^r = \psi((3\ 2\ 2)) = ACTCTC$.*

Now, one can get a condition on linear code over $\mathbb{Z}_4$ that ensures the reversible property in the respective DNA code as given in Lemma 10 as follows.

Lemma 10. *For any generating matrix G with n columns over $\mathbb{Z}_4$, the DNA code $\psi(\langle G\rangle)$ contains reversible DNA strings if and only if the following two properties are satisfied.*

- *The string $(2\ 2\ \dots\ 2)$ is in $\langle G\rangle$.*
- *For each row $(x_1\ x_2\ \dots\ x_n)$ of the matrix G, the reversible string $(x_n\ x_{n-1}\ \dots\ x_1) \in \langle G\rangle$.*

Proof. Consider a G matrix with k rows and n columns over $\mathbb{Z}_4$, where the i^{th} row is $\mathbf{x}_i = (x_{i,1}\ x_{i,2}\ \dots\ x_{i,n})$ for $i = 1, 2, \dots, k$. $\mathbf{y} \Leftrightarrow \sum_{i=1}^k a_i \cdot \mathbf{x}_i \in \langle G\rangle \Leftrightarrow \sum_{i=1}^k a_i \cdot (x_{i,1}\ x_{i,2}\ \dots\ x_{i,n}) \in \langle G\rangle \Leftrightarrow (\sum_{i=1}^k a_i \cdot x_{i,1}\ \ \sum_{i=1}^k a_i \cdot x_{i,2}\ \dots\ \sum_{i=1}^k a_i \cdot x_{i,n}) \in \langle G\rangle \Leftrightarrow (2 + \sum_{i=1}^k a_i \cdot x_{i,n}\ 2 + \sum_{i=1}^k a_i \cdot x_{i,n-1}\ \dots\ 2 + \sum_{i=1}^k a_i \cdot x_{i,1}) \in \langle G\rangle$, where the arguments are obtained from, for any $(z_1\ z_2\ \dots\ z_n) \in \langle G\rangle$, the strings $(z_n\ z_{n-1}\ \dots\ z_1)$ and $(2+z_1\ 2+z_2\ \dots\ 2+z_n)$ are in $\langle G\rangle$ since $(2\ 2\ \dots\ 2) \in \langle G\rangle$. Hence, the result holds. □

Now, define a map $d_\psi : \mathbb{Z}_4 \times \mathbb{Z}_4 \to \mathbb{R}$ *s.t.* $d_\psi(x, y) = d_H(\psi(x), \psi(y))$ for any $x, y \in \mathbb{Z}_4$. For any $\mathbf{x} = (x_1\ x_2\ \dots\ x_n) \in \mathbb{Z}_4^n$ and $\mathbf{y} = (y_1\ y_2\ \dots\ y_n) \in \mathbb{Z}_4^n$, we define $d_\psi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n d_\psi(x_i, y_i)$. Hence, for any $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_4^n$, the distance $d_\psi(\mathbf{x}, \mathbf{y}) = d_H(\psi(\mathbf{x}), \psi(\mathbf{y}))$. For any code $\mathscr{C}$ over $\mathbb{Z}_4$, define $d_\psi = \min\{d_\psi(\mathbf{x}, \mathbf{y}) : \mathbf{x} \neq \mathbf{y}$ and $\mathbf{x}, \mathbf{y} \in \mathscr{C}\}$. The map d_ψ inherits the properties of a distance from d_H, as d_H is itself a distance. Now, one can obtain distance preserving property as given in Lemma 11 as follows.

Lemma 11. *The map $\psi : (\mathbb{Z}_4^n, d_\psi) \to (\mathcal{A}_4^n, d_H)$ is a distance preserving map.*

Proof. For any $\mathbf{x} = (x_1\ x_2\ \dots\ x_n)$ and $\mathbf{y} = (y_1\ y_2\ \dots\ y_n)$ over $\mathbb{Z}_4$, the encoded DNA strings are $\psi(\mathbf{x}) = \psi(x_1)\psi(x_2)\dots\psi(x_n)$ and $\psi(\mathbf{y}) = \psi(y_1)\psi(y_2)\dots\psi(y_n)$ over $\mathcal{A}$. From the distance definition, $d_\psi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n d_H(\psi(x_i), \psi(y_i)) = d_H(\psi(\mathbf{x}), \psi(\mathbf{y}))$. Thus, the map ψ is a distance-preserving map. □

Now, using Lemma 11, one can get the parameter of DNA codes with various properties as given in Theorem 1, Lemma 12, and Lemma 13 as follows.

Theorem 1. *For any (n, M, d_ψ) code $\mathscr{C}$ over $\mathbb{Z}_4$, there is a $(2n, M, d_H)$ DNA code $\psi(\mathscr{C})$ with GC-Content constraint such that each DNA string in $\psi(\mathscr{C})$ and their concatenations are 3 free-structures and 3 free-homopolymers, where $d_\psi = d_H$. In addition, the concatenated DNA strings have GC-Content is again half of their respective lengths.*

Proof. For any (n, M, d) code $\mathscr{C}$ over $\mathbb{Z}_4$, parameters $(2n, M, d_H)$ of DNA code $\psi(\mathscr{C})$ follows from Property 1, bijective property of ψ and Lemma 11. Avoiding secondary structures, GC-Content, and avoiding homopolymers properties for DNA strings and their concatenated strings follows from Proposition 2, Lemma 6 and Lemma 7, and Remark 3. □

Remark 4. *For any (n, M, d_ψ) code over $\mathbb{Z}_4$ in Theorem 1, if $d_\psi \leq n$ then, from Lemma 8, the $(2n, M, d_H)$ DNA code will also satisfy RC constraint.*

Parameters and properties of any DNA codes that are obtained from the union of various DNA codes are discussed in Remark 5, Lemma 12, Remark 6, Lemma 13 as follows.

Remark 5. *For any (n, M, d_ψ) code $\mathscr{C}$ over $\mathbb{Z}_4$, DNA codes $\psi(\mathscr{C})^r$ and $\psi(\mathscr{C}) \cup \psi(\mathscr{C})^r$ are $(2n, M, d_H)$ and is $(2n, M^*, d_H^*)$ code respectively, where $M \leq M^* \leq 2M$, $d_H^* = \min\{d_H, d_H^r\}$ and $d_\psi = d_H$. In addition, if $\mathbf{x}^r \in \psi(\mathscr{C})$ for each $\mathbf{x} \in \psi(\mathscr{C})$ then $M^* = M$ and $d_H^* = d_H$.*

Lemma 12. *For any (n, M, d_ψ) code $\mathscr{C}$ over $\mathbb{Z}_4$, there exists a $(2n, 2M, d_H)$ DNA code $\psi(\mathscr{C}) \cup \psi(\mathscr{C})^c$ with a GC-Content constraint, such that each DNA string in $\psi(\mathscr{C}) \cup \psi(\mathscr{C})^c$ is both 3-free in terms of secondary structures and 3-free in terms of homopolymers, where $d_\psi = d_H$. Furthermore, the concatenated strings of these DNA codewords are also 3-free in terms of homopolymers and maintain a GC-Content equal to half of their respective lengths.*

Proof. The parameter of the code $\psi(\mathscr{C}) \cup \psi(\mathscr{C})^c$ is derived from Lemma 9. Properties pertaining to secondary structures and homopolymers for DNA strings in $\psi(\mathscr{C}) \cup \psi(\mathscr{C})^c$ are established by Proposition 2 and Lemma 7, respectively. Constraints related to reverse complementarity and GC-Content are demonstrated in Lemma 5 and Lemma 6. Finally, the concatenation property is a consequence of Remark 3 and Remark 9. □

Remark 6. *For any code $\mathscr{C}$ over $\mathbb{Z}_4$ in Lemma 12, if $d_\psi \leq n$ then the DNA code $\psi(\mathscr{C}) \cup \psi(\mathscr{C})^c$ will also satisfy RC constraint.*

Lemma 13. *For any (n, M, d_ψ) code $\mathscr{C}$ over $\mathbb{Z}_4$, there is a $(2n, M^*, d_H^*)$ DNA code $\psi(\mathscr{C}) \cup \psi(\mathscr{C})^c \cup \psi(\mathscr{C})^r \cup \psi(\mathscr{C})^{rc}$ with R, RC and GC-Content constraints, and the DNA codewords*

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

are 3 *free-structures and* 3 *free-homopolymers strings, where* $M \leq M^* \leq 2M$, $d_\psi \geq d_H^*$. *Also, concatenated strings of those DNA codewords are* 3 *free-homopolymers and have* GC*-Content half of their lengths. Also, if* $\mathbf{x}^r \in \psi(\mathscr{C})$ *for each* $\mathbf{x} \in \psi(\mathscr{C})$ *then* $M^* = M$ *and* $d_H^* = d_H$.

Proof. For any DNA string $\mathbf{x}$, $\mathbf{x}^{rc} = (\mathbf{x}^r)^c = (\mathbf{x}^c)^r$, and therefore, for any DNA code $\mathscr{C}_{DNA}$, $\mathscr{C}_{DNA} \cup \mathscr{C}_{DNA}^c \cup \mathscr{C}_{DNA}^r \cup \mathscr{C}_{DNA}^{rc} = (\mathscr{C}_{DNA} \cup \mathscr{C}_{DNA}^c) \cup (\mathscr{C}_{DNA}^r \cup \mathscr{C}_{DNA}^c)^r$. The parameter of the DNA code $\psi(\mathscr{C}) \cup \psi(\mathscr{C})^c \cup \psi(\mathscr{C})^r \cup \psi(\mathscr{C})^{rc}$ (= $\mathscr{C}_{DNA}^*$) follows from Lemma 9 and Remark 5. Note that, for any DNA code $\mathscr{C}_{DNA}$ over $\mathcal{A}_4$, the DNA code $\mathscr{C}_{DNA}^r$ is also defined over $\mathcal{A}_4$, and thus, properties on avoiding secondary structures and homopolymers for DNA strings in $\mathscr{C}_{DNA}^*$ are followed from Proposition 2 and Lemma 7. In addition, the RC and GC-Content constraints for these DNA strings are followed from Lemma 8 and Lemma 6. Further, the concatenation property follows from Remark 3 and Remark 5. □

V. Examples of DNA Codes from Linear Codes over $\mathbb{Z}_4$

This section presents a comprehensive analysis of DNA codes derived from linear codes over $\mathbb{Z}_4$, demonstrating the practical application of our theoretical framework and highlighting the properties of the resulting DNA codes. We consider six linear codes over $\mathbb{Z}_4$ with generating matrices G_i ($i = 0, 1, 2, 3, 4, 5$). The parameters of the corresponding DNA codes $\psi(\langle G_i \rangle)$ are summarized in Table II. The generating matrices are defined as follows:

$$G_0 = (2\ 2), \qquad G_1 = (1\ 1\ 3\ 3),$$

$$G_2 = \begin{pmatrix} 1 & 0 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{pmatrix}, \qquad G_3 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 2 & 2 \end{pmatrix},$$

$$G_4 = \begin{pmatrix} 1 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 1 & 2 & 1 \end{pmatrix}, \quad G_5 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 & 2 & 2 \end{pmatrix}.$$

Our analysis reveals that the DNA codes $\psi(\langle G_i \rangle)$, $\psi(\langle G_i \rangle) \cup \psi(\langle G_i \rangle)^c$ and $\psi(\langle G_i \rangle) \cup \psi(\langle G_i \rangle)^c \cup \psi(\langle G_i \rangle)^r \cup \psi(\langle G_i \rangle)^{rc}$ satisfy RC and GC-Content constraints. In addition, each codeword in these DNA codes $\psi(\langle G_i \rangle)$, $\psi(\langle G_i \rangle) \cup \psi(\langle G_i \rangle)^c$ and $\psi(\langle G_i \rangle) \cup \psi(\langle G_i \rangle)^c \cup \psi(\langle G_i \rangle)^r \cup \psi(\langle G_i \rangle)^{rc}$ is 3 free-structures and also 3 free-homopolymers. Any DNA string obtained by concatenating codewords of DNA codes $\psi(\langle G_i \rangle)$ is again 3 free-structures, and 3 free-homopolymers and also has GC-contain half of its length. Any DNA string obtained by concatenating codewords of DNA codes $\psi(\langle G_i \rangle) \cup \psi(\langle G_i \rangle)^c$ and $\psi(\langle G_i \rangle) \cup \psi(\langle G_i \rangle)^c \cup \psi(\langle G_i \rangle)^r \cup \psi(\langle G_i \rangle)^{rc}$ is 3 free-homopolymers, and also has GC-contain half of its length. Any DNA code $\psi(\langle G_i \rangle) \cup \psi(\langle G_i \rangle)^c$ holds R constraint. For $i = 0, 1, 2, 3$, DNA codes $\psi(\langle G_i \rangle)$ and $\psi(\langle G_i \rangle) \cup \psi(\langle G_i \rangle)^c$ satisfy R constraint since G_i holds the properties given in Lemma 10. These examples demonstrate the effectiveness of our approach in generating DNA codes with desirable properties, balancing error-correction capabilities, structural stability, and biological compatibility for various applications in DNA-based information systems.

VI. Families of DNA codes

Motivated by the Simplex code and Reed-Muller code, families of DNA codes are constructed in this section.

A. Modified Simplex DNA Codes of Type 1

Motivated by [62, Chapter 3], for a given positive integer k, consider a linear code $\mathscr{C}_{k+1}^\alpha$ over $\mathbb{Z}_4$ with generating matrix $G_{k+1} = \begin{pmatrix} 1\ 1\ \ldots\ 1 \\ G_k^\alpha \end{pmatrix}$, where $G_1^\alpha = (0\ 1\ 2\ 3)$, and, for $k \geq 2$, the matrix G_k^α is defined inductively as

$$G_k^\alpha = \left(\begin{array}{c|c|c|c} 0\ 0\ \ldots\ 0 & 1\ 1\ \ldots\ 1 & 2\ 2\ \ldots\ 2 & 3\ 3\ \ldots\ 3 \\ \hline G_{k-1}^\alpha & G_{k-1}^\alpha & G_{k-1}^\alpha & G_{k-1}^\alpha \end{array} \right).$$

The parameters of $\mathscr{C}_{k+1}^\alpha$ and $\psi(\mathscr{C}_{k+1}^\alpha)$ are discussed in Lemma 14 and Lemma 15 as follows.

Lemma 14. *For any positive integer* k, $\mathscr{C}_{k+1}^\alpha$ *is a quaternary code of length* 2^{2k}, *size* 2^{2k+2}, *the MHD* 2^{2k-1}, *and the minimum distance* $d_\psi = 2^{2k}$.

Proof. From the symmetry of the generating matrix G_{k+1} for $\mathscr{C}_{k+1}^\alpha$, it contains 4^k columns, and the result on length holds. The generating matrix G_{k+1} is of type $\{k+1, 0\}$ over $\mathbb{Z}_4$. Therefore, the size of the code is 4^{k+1}, and the result on code size holds. Now, from computation, one can find the MHD and the minimum distance for the set $\langle \begin{pmatrix} \mathbf{i}_4 \\ G_2 \end{pmatrix} \rangle$ are 2 and 4 respectively for each $i = 0, 1, 2, 3$, where $\mathbf{i}_4 = (i\ i\ i\ i)$. In the matrix G_{k+1}, there are 4^{k-1} disjoint blocks each containing four columns, where two blocks are disjoint if they do not share a common column. Therefore, for the code $\mathscr{C}_{k+1}^\alpha$, the MHD $d_H \geq 2 \cdot 4^{k-1}$ and the minimum distance $d_\psi \geq 4 \cdot 4^{k-1} = 2^{2k}$. But, for $\mathbf{0}_{2^{2k}}, (0\ 2\ 0\ 2\ \ldots\ 0\ 2) \in \langle G_{k+1} \rangle$, the Hamming distance $d_H(\mathbf{0}_{2^{2k}}, (0\ 2\ 0\ 2\ \ldots\ 0\ 2)) = 2^{2k-1}$ and the distance $d_\psi(\mathbf{0}_{2^{2k}}, (0\ 2\ 0\ 2\ \ldots\ 0\ 2)) = 4 \cdot 4^{k-1}$. The result holds for the MHD. □

Lemma 15. *For any positive integer* k, *the code* $\psi(\mathscr{C}_{k+1}^\alpha)$ *is an* $(2^{2k+1}, 2^{2k+2}, 2^{2k})$ *DNA code.*

Proof. The parameter of the code $\mathscr{C}_{k+1}^\alpha$ is given in Lemma 14, and therefore, from Theorem 1, one can obtain the DNA code $\psi(\mathscr{C}_{k+1}^\alpha)$ parameters. □

Remark 7. *The DNA code* $\psi(\mathscr{C}_{k+1}^\alpha)$ *holds, R, RC, and* GC*-Content constraints. The DNA codewords are also* 3 *free-homopolymers and* 3 *free-structures. Furthermore, the DNA strings formed by concatenating these codewords satisfy all of the properties.*

Remark 8. *For the DNA code* $\psi(\mathscr{C}_{k+1}^\alpha)$, *the code rate is* $\frac{k+1}{2^{2k+1}}$ ($\leq \frac{1}{4}$), *and the RMHD is* $\frac{1}{2}$. *Also note the asymptotic RHMD* $\lim_{n\to\infty} \left(\frac{d_H}{n}\right)$ *is* $\frac{1}{2}$. *For* $\psi(\mathscr{C}_2^\alpha)$, *the code rate is* $\frac{1}{4}$.

B. Modified Simplex DNA Codes of Type 2

Motivated by [62, Chapter 3], for a given positive integer k, consider a linear code $\mathscr{C}_{k+1}^\alpha$ over $\mathbb{Z}_4$ with generating matrix

$$G_{k+1} = \begin{pmatrix} \mathbf{1}_{4^k} \\ \mathbf{1}_{2^{2k-2}}\ \mathbf{2}_{2^{2k-1}}\ \mathbf{3}_{2^{2k-2}} \\ G_k^\alpha \end{pmatrix},$$

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

TABLE II
SOME EXAMPLES OF DNA CODES ARE LISTED IN THE TABLE, WHERE THOSE DNA CODES ARE OBTAINED FROM LINEAR CODES OVER $\mathbb{Z}_4$, WHERE $\psi(\langle G_i\rangle)^* = \psi(\langle G_i\rangle) \cup \psi(\langle G_i\rangle)^c$ AND $\psi(\langle G_i\rangle)^{**} = \psi(\langle G_i\rangle) \cup \psi(\langle G_i\rangle)^c \cup \psi(\langle G_i\rangle)^r \cup \psi(\langle G_i\rangle)^{rc}$.

Generating matrix G_i	parameter of $\psi(\langle G_i\rangle)$	parameter of $\psi(\langle G_i\rangle)^*$	parameter of $\psi(\langle G_i\rangle)^{**}$	R constraint for $\psi(\langle G_i\rangle)$ and $\psi(\langle G_i\rangle)^*$
G_0	$(4, 2, 4)$	$(4, 4, 4)$	$(4, 4, 4)$	satisfy
G_1	$(8, 4, 6)$	$(8, 8, 6)$	$(8, 8, 6)$	satisfy
G_2	$(8, 16, 4)$	$(8, 32, 4)$	$(8, 32, 4)$	satisfy
G_3	$(8, 16, 4)$	$(8, 32, 4)$	$(8, 32, 4)$	satisfy
G_4	$(10, 64, 3)$	$(10, 128, 3)$	$(10, 240, 1)$	do not satisfy
G_5	$(12, 128, 3)$	$(12, 256, 3)$	$(12, 480, 2)$	do not satisfy

where $G_1^\alpha = (0\ 1\ 2\ 3)$, and, for $k \geq 2$, the matrix G_k^α is defined inductively as

$$G_k^\alpha = \left(\begin{array}{c|c|c|c} 0\ 0\ \ldots\ 0 & 1\ 1\ \ldots\ 1 & 2\ 2\ \ldots\ 2 & 3\ 3\ \ldots\ 3 \\ \hline \bar{G}_{k-1}^\alpha & \bar{G}_{k-1}^\alpha & \bar{G}_{k-1}^\alpha & \bar{G}_{k-1}^\alpha \end{array}\right).$$

The parameters of $\mathscr{C}_{k+2}^{2\alpha}$ and $\psi(\mathscr{C}_{k+2}^{2\alpha})$ are discussed in Lemma 16 and Lemma 17 as follows.

Lemma 16. *For any positive integer k, $\mathscr{C}_{k+2}^{2\alpha}$ is a quaternary code of length 2^{2k}, size 2^{2k+4}, the MHD 2^{2k-1}, and the minimum distance $d_\psi = 2^{2k-1}$.*

Proof. The proof follows a similar argument to that of Lemma 14, with the additional consideration that both the minimum Hamming distance (MHD) and the minimum distance for the code generated by $\langle \begin{pmatrix} \mathbf{i}_4 \\ G_2 \end{pmatrix} \rangle$ are identical and equal to 2. □

Lemma 17. *For any positive integer k, the code $\psi(\mathscr{C}_{k+2}^{2\alpha})$ is an $(2^{2k+1}, 2^{2k+4}, 2^{2k-1})$ DNA code.*

Proof. The result holds from Theorem 1 and Lemma 16. □

Remark 9. *The DNA code $\psi(\mathscr{C}_{k+2}^{2\alpha})$ holds, R, RC, and GC-Content constraints. Further, each DNA codeword is 3 free-homopolymers and 3 free-structures. In addition, the DNA strings obtained by concatenating these DNA codewords again satisfy all the properties.*

Remark 10. *For the DNA code $\psi(\mathscr{C}_{k+2}^{2\alpha})$, the code rate is $\frac{k+2}{2^{2k+1}}$ ($\leq \frac{1}{8}\log_4 32 = 0.3125$), and the RMHD is $\frac{1}{4}$. The asymptotic RHMD $\lim_{n\to\infty}\left(\frac{d_H}{n}\right)$ is $\frac{1}{4}$. The code rate of $\psi(\mathscr{C}_3^{2\alpha})$ is 0.3750.*

C. Modified Hamming DNA Codes

Motivated from the relation between binary Hamming code and binary Simplex code, in this section, for any positive integer k, we proposed Hamming type code obtained from G_k^α as given in Section VI-A. Consider G_k' = $[I_k\ A]$ is the matrix obtained from G_k^α by deleting all zero column. Now, compute G_k^H = $[I_{4^k-k-1}\ A^t]$, where A^t is the transpose of the matrix.

Lemma 18. *For any positive integer k, $\mathscr{C}_k^H$ is a quaternary code of length 2^{2k}, size $2^{(2^{2k+1}-2k-2)}$, the MHD 3, and the minimum distance d_ψ = 3.*

Proof. The proof for length and size is similar to the proof of Lemma 14. For the Hamming distance d_H' of the Quaternary codes, observe the fact the $d_\psi \geq d_H' = 3$, and thus, the MHD for the DNA code is at least 3. But, one can find two rows of the generating matrix G_k^H with only two positions are non-zero element such that the distance d_ψ is 3, and thus the Hamming distance of the DNA code is 3. □

Lemma 19. *For any positive integer k, the code $\psi(\mathscr{C}_k^H)$ is an $(2^{2k+1}, 2^{(2^{2k+1}-2k-2)}, 3)$ DNA code.*

Proof. The result holds from Theorem 1 and Lemma 16. □

Remark 11. *The DNA code $\psi(\mathscr{C}_k^H)$ holds, RC, and GC-Content constraints. Further, the DNA codewords are 3 free-homopolymers and also 3 free-structures. In addition, the DNA strings obtained by concatenating these DNA codewords again satisfy all the properties.*

Remark 12. *For the DNA code $\psi(\mathscr{C}_k^H)$, the code rate is $\frac{2^{2k}-k-1}{2^{2k+1}}$, and the RMHD is 3. Also, the code rate of the DNA code $\psi(\mathscr{C}_k^H)$ approaches to $\frac{1}{2}$ as the length approaches to infinity.*

D. Reed-Muller Type DNA Codes

Let m and r be integers such that $0 \leq r \leq m$. The generator matrix for the (r, m)-th order Reed-Muller type code, denoted as $\mathscr{R}(r, m)$ over $\mathbb{Z}_4$, is defined recursively as follows:

$$G_{r,m} = \begin{pmatrix} G_{r,m-1} & G_{r,m-1} \\ \mathbf{0} & G_{r-1,m-1} \end{pmatrix} \qquad \text{for } 1 \leq r \leq m-1,$$

where $G_{m,m} = \begin{pmatrix} G_{m-1,m} \\ \mathbf{0}\ 1\ 2\ 3 \end{pmatrix}$ and $G_{0,m}$ is a single row vector of length 2^{m+1} populated entirely with ones. Here, $\mathbf{0}$ denotes a zero matrix of appropriate dimensions.

Theorem 2. *Let m and r be integers such that $0 \leq r \leq m$. The code $\psi(\mathscr{R}(r, m))$ is the DNA code with length 2^{m+2}, size $4^{\sum_{i=0}^{r}\binom{m}{i}}$, and the MHD d_H = 2^{m-r-1} for $r < m$, and d_H = 4 for $r = m$.*

Proof. The length of the DNA code $\psi(\mathscr{R}(r, m))$ can be established by induction on the parameter m. Specifically, the length is given by 2^{m+2}, with the base case corresponding to $\psi(\mathscr{R}(1, 1))$.

Regarding the code size, note that the generator matrix $G_{r,m}$ has type $\{\sum_{i=0}^{r}\binom{m}{i}, 0\}$. Consequently, applying Theorem 1 yields that the cardinality of the DNA code $\psi(\mathscr{R}(r, m))$ is $4^{\sum_{i=0}^{r}\binom{m}{i}}$.

□

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

Remark 13. *The DNA code $\psi(\mathscr{R}(r,m))$ holds R, RC and GC-Content constraints. Further, the DNA codewords are 3 free-homopolymers and also 3 free-structures. In addition, any concatenation of these DNA codewords results in 3 free-homopolymers, 3 free-structures and GC-Content is half of its length.*

Remark 14. *For the DNA code $\psi(\mathscr{R}(r,m))$, the code rate is $\frac{\sum_{i=0}^{r}\binom{m}{i}}{2^{m+2}}$ ($\leq \frac{1}{2}$). For $r < m$, the RMHD is $\frac{2^{m-r-1}}{2^{m+2}} = 2^{1-r}$, and, for $r = m$, the RMHD is $\frac{4}{2^{m+2}}$.*

Remark 15. *For given positive integer Δ ($1 \leq \Delta < \lfloor m/2 \rfloor$), the code rate $\lim_{n\to\infty}\left(\frac{1}{n}\log_4 M\right)$ of the DNA code $\psi(\mathscr{R}(r,m))$ is approaching $\frac{1}{2}$ for large lengths, where $\Delta = m - r$.*

VII. Discussions and Comparisons

This section compares the properties of our constructed DNA codes with those reported in the literature and discusses their characteristics.

A. Code Rate Trade-Off and Secondary Structures

As summarized in Table I, increasing the word length parameter t generally yields higher code rates due to the expansion of the admissible codeword set. However, this increase comes at the expense of substantially greater computational complexity during code construction, as noted in Remark 1. Consequently, the choice of t must strike a balance between improving code rates and practical computational resource constraints.

Figure 3 illustrates this trade-off by plotting the code rate against code length $n = 1, 2, \ldots, 11$ for DNA codes without secondary structures having minimum stem length $t = 1, 2, 3, 4, 5$. For a fixed code length n, the code rate increases with t and ultimately approaches one when $t > \lfloor n/2 \rfloor$.

B. Maximum Code Rates under Constraints

[63] and [64] present closed-form Sphere Packing (SP) and Gilbert–Varshamov (GV) bounds for DNA codes with GC-content and without homopolymers, respectively. Figure 5 compares these bounds for general codes, codes without homopolymers, and codes with GC-content, alongside the Modified Hamming DNA codes rate proposed here. Unlike the SP and GV bounds that consider single constraints, the Modified Hamming codes satisfy multiple constraints, including Homopolymer and GC-content constraints. The gap in code rate reflects this difference in constraint enforcement.

Figure 4 compares the code rate versus code length for DNA codes designed under multiple biologically relevant constraints (i) GC-content balancing, (ii) Homopolymer runs limited to length two, (iii) Avoidance of secondary structures with stem length ≥ 3, (iv) Combinations of the above constraints. The labels *Sec Str* and *Homo* represent the absence of secondary structures with stem length exceeding two and homopolymer runs restricted to at most two nucleotides, respectively. The plotted curves serve as upper bounds on the achievable code rate under each respective constraint. Bounds were estimated by enumerating all 4^n DNA sequences of length n using MATLAB, filtering those that violate constraints, and plotting the results in Figure 4. All considered DNA codes satisfy the stated restrictions, although note that concatenation properties related to secondary structures are not incorporated in these results. The code rate gap between the bound in Figure 4 and the constructed DNA codes is due to the concatenation property and error correction property in the constructed DNA codes.

C. Comparison with Existing Literature

Previous works have addressed subsets of the constraints considered here, but the present work incorporates a broader set of biologically motivated constraints simultaneously. For example:

- In [58], the authors studied substitution error mitigation via unconstrained and constrained DNA codes, focusing primarily on avoiding homopolymers and controlling GC-content. By contrast, our constructions additionally enforce secondary structure avoidance and simultaneously handle R and RC constraints.
- The work in [65] constructs DNA codes that avoid secondary structures by eliminating reverse-complement subsequences. Our approach generalizes this by incorporating additional considerations including G-T pairing, alongside canonical A-T and G-C pairings.
- In [66], DNA codes ensure R, RC, and GC-content constraints while avoiding homopolymers and secondary structures, with code lengths restricted to multiples of three. Unlike this, our constructions support code lengths that are powers of two, thereby offering increased flexibility for DNA data encoding.

D. Comparison and Analysis of DNA Codes

The properties of the obtained DNA codes are compared in Table III with the properties of DNA codes reported in the literature. For DNA codes listed in Table III, the term Yes* stands for the concatenated DNA strings that are 3 free-homopolymers and the GC-Content of the concatenated DNA strings is half of their respective lengths. Note that concatenated DNA strings of those DNA codes marked with Yes* in Table III are 3-free-homopolymers but not 3 free-structures.

Note that secondary structures avoiding DNA codes with RC constraint are obtained from codes over the alphabet $\{AA, AC, CA, CC, TC\}$ in [67], and similarly, secondary structures and homopolymers avoiding and GC balanced DNA codes with R and RC constraints are obtained from codes over the alphabet of size 6 in [66].

E. Properties of DNA Codes

The alphabet sets considered in [66], [67] are the subsets of the sets $\mathscr{A}$ with $t = 2$ and $t = 3$ in Table I, respectively. Comparisons of properties of DNA codes constructed in this paper and available in the literature are made in Table III. All DNA codes listed from the literature in Table III are single error-correcting codes. For example, each DNA

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

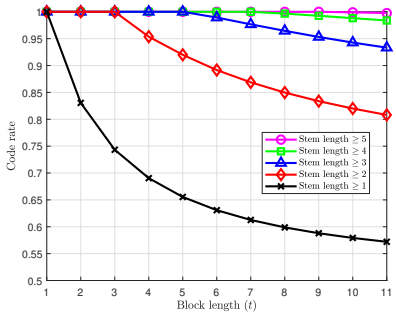

Fig. 3. Trade-off between code rate and length n for DNA codes without secondary structures of stem length more than and equal to $\ell = 1, 2, 3, 4, 5$ and $t = 1, 2, \ldots, 11$.

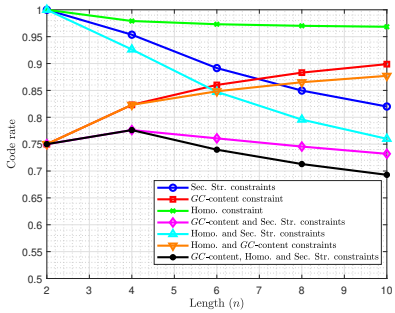

Fig. 4. Plot of code rate versus code length n ($n = 2, 4, \ldots, 10$) for DNA codes with GC-content constraint and avoidance of secondary structures of stem length more than two and homopolymers of runlength more than two.

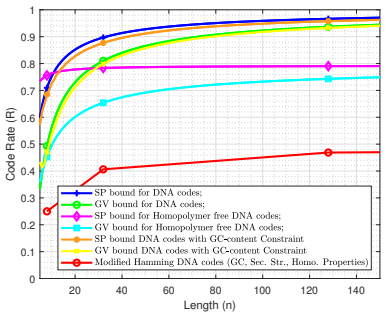

Fig. 5. For $d_H = 3$, the code rate for the modified Hamming DNA code is plotted with the code rate obtained from Sphere Packing (SP) and Gilbert–Varshamov (GV) bounds for DNA code in general, DNA codes without Homopolymers and DNA codes with GC-content constraint.

TABLE III
COMPARISONS OF PROPERTIES OF DNA CODES.

DNA Codes (n, M, d_H)	Code Rate $\frac{\log_4 M}{n}$	Relative min Hamming Distance $\frac{d_H}{n}$	holds Concatenation Property	ℓ free-structures Property	ℓ free-homopolymers Property	holds RC Constraint	holds R Constraint	holds GC-Content Constraint
$(15, 15, 12)$ code, Example 9 [68]	0.1494	0.8000	No	No	$\ell = 5$	No	Yes	Yes
$(5, 6, 4)$ code, Example 2 [15]	0.2585	0.8000	No	$\ell = 4$	$\ell = 5$	No	Yes	No
$f(\psi(\mathcal{H}_2))$ [67]	0.3870	0.2500	Yes	$\ell = 3$	$\ell = 5$	Yes	No	No
$(7, 49, 4)$ code, Example 4 [15]	0.4011	0.5714	No	$\ell = 3$	$\ell = 5$	No	Yes	No
$\mathscr{C} \cup \mathscr{C}^c$, $\mathscr{C} = f(\psi(\mathcal{H}_2))$ [67]	0.4287	0.2500	No	$\ell = 3$	$\ell = 5$	Yes	No	No
$(8, 224, 4)$ code, Table III [21]	0.4880	0.5000	No	No	No	Yes	No	Yes
$(8, 256, 4)$ code, Table III [21]	0.5000	0.5000	No	No	No	Yes	Yes	No
$\psi(\langle G_4 \rangle)^*$ Table II	0.3500	0.3000	Yes*	$\ell = 3$	$\ell = 3$	Yes	No	Yes
$\psi(\langle G_5 \rangle)^*$ Table II	0.3333	0.2500	Yes*	$\ell = 3$	$\ell = 3$	Yes	No	Yes
$\psi(\langle G_2 \rangle)^{**}$ Table II	0.3125	0.5000	Yes*	$\ell = 3$	$\ell = 3$	Yes	Yes	Yes
$\psi(\langle G_3 \rangle)^{**}$ Table II	0.3125	0.5000	Yes*	$\ell = 3$	$\ell = 3$	Yes	Yes	Yes
$\psi(\langle G_4 \rangle)$ Table II	0.3000	0.3000	Yes	$\ell = 3$	$\ell = 3$	Yes	Yes	Yes
$\psi(\langle G_0 \rangle)^*$ Table II	0.2500	1.0000	Yes*	$\ell = 3$	$\ell = 2$	Yes	No	Yes
$\psi(\langle G_2 \rangle)$ Table II	0.2500	0.5000	Yes	$\ell = 3$	$\ell = 3$	Yes	Yes	Yes
$\psi(\mathscr{R}(1, 1))$	0.2500	0.5000	Yes	$\ell = 3$	$\ell = 3$	Yes	Yes	Yes
$\psi(\mathscr{C}_2^{\alpha})$	0.2500	0.5000	Yes	$\ell = 3$	$\ell = 3$	Yes	Yes	Yes
$\psi(\langle G_1 \rangle)^{**}$ Table II	0.1875	0.7500	Yes*	$\ell = 3$	$\ell = 3$	Yes	Yes	Yes

string in $\psi(\mathscr{R}(r, m))$ $(r, m \leq 4)$ is 3 free-homopolymers, and secondary structures are not exhibited in Vienna [18] and The mfold Web Server [17]. In [15, Example 2 and Example 4], DNA codes are designed that avoid secondary structures, but homopolymer property is not considered in [15]. In [15, Example 2], the DNA code is obtained with length 5, $d_H = d_H^r = 4$ and $d_H^{rc} = 3$. In [15, Example 4], [68, Example 9], and $(8, 224, 4)$ DNA code (Table III in [21]), DNA codes satisfy GC-Content constraint. The $(8, 256, 4)$ DNA code holds the R constraint in [21, Table III]. All DNA strings in $(15, 15, 12)$ code as given in [68, Example 9] has identical GC-Content and it is equal to 8. The DNA strings in DNA codes $f(\psi(\mathscr{C}_2)) \cup f(\psi(\mathscr{C}_2))^c$, $f(\psi(\mathcal{H}_2))$, and $f(\psi(\mathcal{H}_2)) \cup f(\psi(\mathcal{H}_2))^c$ are 5 free-homopolymers [67].

F. DNA Codes and Bounds

One can observe that the $(4, 4, 4)$ DNA code as given in Table II holds the Singleton bound for quaternary codes. Also, DNA codes of parameters $(4, 4, 4)$ and $(8, 8, 6)$ achieve the Gilbert-Varshamov bound for non-linear quaternary codes. Therefore, $(4, 4, 4)$ and $(8, 8, 6)$ DNA codes are optimal codes that satisfy R, RC, and GC-constraint constraints. In addition, the DNA strings are 3 free-structures and also 3 free-homopolymers. The $(4, 4, 4)$ DNA code is $\{CTCT, TCTC, GAGA, AGAG\}$ and the $(8, 8, 6)$ DNA code is $\{CACAACAC,\ ACACCACA,\ CTCTCTCT,\ TCTCTCTC,\ GTGTTGTG,\ TGTGGTGT,\ GAGAGAGA,\ AGAGAGAG\}$.

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

G. Code Rate and relative minimum Hamming Distance

The families of DNA codes obtained from Modified Simplex Codes of Type 1 and of Type 2, and Reed-Muller Type codes have either a non-vanishing code rate or a non-vanishing relative minimum Hamming distance for large lengths. Note that, for Modified Simplex DNA Codes of Type 1 and of Type 2, the RMHD is approaching $1/2$ (see Remark 8) and $1/4$ (see Remark 10) for large lengths, respectively. Also, for given $\Delta = m - r$, the code rate is approaching $1/2$ for DNA codes $\psi(\mathscr{R}(r, m))$ for large lengths (see Remark 15).

REFERENCES

[1] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, p. 77, 2013.

[3] O. Milenkovic and C. Pan, "DNA-Based data storage systems: A review of implementations and code constructions," *IEEE Transactions on Communications*, vol. 72, no. 7, pp. 3803–3828, 2024.

[4] O. Sabary, H. M. Kiah, P. H. Siegel, and E. Yaakobi, "Survey for a decade of coding for DNA storage," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 10, no. 2, pp. 253–271, 2024.

[5] A. Doricchi, C. M. Platnich, A. Gimpel, F. Horn, M. Earle, G. Lanzavecchia, A. L. Cortajarena, L. M. Liz-Marzán, N. Liu, R. Heckel, R. N. Grass, R. Krahne, U. F. Keyser, and D. Garoli, "Emerging approaches to DNA data storage: Challenges and prospects," *ACS Nano*, vol. 16, no. 11, pp. 17 552–17 571, Nov 2022.

[6] P. Y. De Silva and G. U. Ganegoda, "New trends of digital data storage in DNA," *Biomed Res Int*, vol. 2016, p. 8072463, Sep. 2016.

[7] T. Buko, N. Tuczko, and T. Ishikawa, "DNA data storage," *BioTech*, vol. 12, no. 2, p. 44, Jun 2023.

[8] Y. Dong, F. Sun, Z. Ping, Q. Ouyang, and L. Qian, "DNA storage: research landscape and future prospects," *National Science Review*, vol. 7, no. 6, pp. 1092–1107, 01 2020.

[9] T. Heinis, R. Sokolovskii, and J. J. Alnasir, "Survey of information encoding techniques for DNA," *ACM Comput. Surv.*, vol. 56, no. 4, Nov 2023.

[10] S. M. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 1, no. 3, pp. 230–248, 2015.

[11] A. Marathe, A. E. Condon, and R. M. Corn, "On combinatorial DNA word design," *Journal of Computational Biology*, vol. 8, no. 3, pp. 201–219, 2001.

[12] K. A. S. Immink and K. Cai, "Properties and constructions of constrained codes for DNA-based data storage," *IEEE Access*, vol. 8, pp. 49 523–49 531, 2020.

[13] T. T. Nguyen, K. Cai, K. A. Schouhamer Immink, and H. M. Kiah, "Capacity-approaching constrained codes with error correction for DNA-based data storage," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 5602–5613, 2021.

[14] S. M. H. Tabatabaei Yazdi, H. M. Kiah, R. Gabrys, and O. Milenkovic, "Mutually uncorrelated primers for DNA-based data storage," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6283–6296, 2018.

[15] O. Milenkovic and N. Kashyap, "On the design of codes for DNA computing," in *Coding and Cryptography*, Ø. Ytrehus, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 100–119.

[16] A. Peselis and A. Serganov, "Structure and function of pseudoknots involved in gene expression control," *Wiley Interdiscip Rev RNA*, vol. 5, no. 6, pp. 803–822, 2014.

[17] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3406–3415, 2003.

[18] A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker, "The Vienna RNA Websuite," *Nucleic Acids Research*, vol. 36, no. suppl_2, pp. W70 – W74, March 2008.

[19] O. Milenkovic, "Constrained coding for context-free languages with applications to genetic sequence modelling," in *IEEE International Symposium on Information Theory*, 2007, pp. 1686–1690.

[20] O. Milenkovic and E. Soljanin, "Enumeration of RNA secondary structures: A constrained coding approach," in *Fortieth Asilomar Conference on Signals, Systems and Computers*, October 2006, pp. 1954–1958.

[21] D. Limbachiya, K. G. Benerjee, B. Rao, and M. K. Gupta, "On DNA codes using the ring $\mathbb{Z}_4 + w\mathbb{Z}_4$," in *IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 2401–2405.

[22] V. V. Rykov, A. J. Macula, D. C. Torney, and P. S. White, "DNA sequences and quaternary cyclic codes," in *IEEE International Symposium on Information Theory (ISIT)*, 2001, pp. 248–248.

[23] S. Das, K. G. Benerjee, and A. Banerjee, "On DNA codes over the non-chain ring $\mathbb{Z}_4 + u\mathbb{Z}_4 + u^2\mathbb{Z}_4$ with $u^3 = 1$," in *IEEE Information Theory Workshop (ITW)*, 2022, pp. 660–665.

[24] Y. M. Chee, H. M. Kiah, and H. Wei, "Efficient and explicit balanced primer codes," *IEEE Transactions on Information Theory*, vol. 66, no. 9, pp. 5344–5357, 2020.

[25] L. Deng, Y. Wang, M. Noor-A-Rahim, Y. L. Guan, Z. Shi, E. Gunawan, and C. L. Poh, "Optimized code design for constrained DNA data storage with asymmetric errors," *IEEE Access*, vol. 7, pp. 84 107–84 121, 2019.

[26] Y. Liu, X. He, and X. Tang, "Capacity-achieving constrained codes with GC-content and runlength limits for DNA storage," in *IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 198–203.

[27] J. H. Weber, J. A. M. De Groot, and C. J. Van Leeuwen, "On single-error-detecting codes for DNA-based data storage," *IEEE Communications Letters*, pp. 1–1, 2020.

[28] P. Gaborit and O. D. King, "Linear constructions for DNA codes," *Theoretical Computer Science*, vol. 334, no. 1, pp. 99 – 113, 2005.

[29] D. Tulpan, D. H. Smith, and R. Montemanni, "Thermodynamic post-processing versus GC-content pre-processing for DNA codes satisfying the Hamming distance and reverse-complement constraints," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 2, pp. 441–452, 2014.

[30] N. Aboluion, D. H. Smith, and S. Perkins, "Linear and nonlinear constructions of DNA codes with Hamming distance d, constant GC-content and a reverse-complement constraint," *Discrete Mathematics*, vol. 312, no. 5, pp. 1062 – 1075, 2012.

[31] K. G. Benerjee, S. Deb, and M. K. Gupta, "On conflict free DNA codes," *Cryptography and Communications*, vol. 13, no. 1, pp. 143–171, 2021.

[32] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2331–2351, 2020.

[33] T. Thanh Nguyen, K. Cai, W. Song, and K. A. Schouhamer Immink, "Optimal single chromosome-inversion correcting codes for data storage in live DNA," in *IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 1791–1796.

[34] H. Wei and M. Schwartz, "Improved coding over sets for DNA-based data storage," *IEEE Transactions on Information Theory*, vol. 68, no. 1, pp. 118–129, 2022.

[35] W. Zhang, Z. Chen, and Z. Wang, "Limited-magnitude error correction for probability vectors in DNA storage," in *IEEE International Conference on Communications (ICC)*, 2022, pp. 3460–3465.

[36] T. Chen, Y. Ma, and X. Zhang, "Optimal codes with small constant weight in ℓ_1-metric," *IEEE Transactions on Information Theory*, vol. 67, no. 7, pp. 4239–4254, 2021.

[37] M. Cheraghchi, R. Gabrys, O. Milenkovic, and J. Ribeiro, "Coded trace reconstruction," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6084–6103, 2020.

[38] T. Thanh Nguyen, K. Cai, and P. H. Siegel, "A new version of q-ary varshamov-tenengolts codes with more efficient encoders: The differential VT codes and the differential shifted VT codes," *IEEE Transactions on Information Theory*, vol. 70, no. 10, pp. 6989–7004, 2024.

[39] Y. Yehezkeally, D. Bar-Lev, S. Marcovich, and E. Yaakobi, "Generalized unique reconstruction from substrings," *IEEE Transactions on Information Theory*, vol. 69, no. 9, pp. 5648–5659, 2023.

[40] D. Goshkoder, N. Polyanskii, and I. Vorobyev, "Codes correcting long duplication errors," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 10, no. 2, pp. 272–288, 2024.

[41] R. Gabrys, E. Yaakobi, and O. Milenkovic, "Codes in the damerau distance for DNA storage," in *IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 2644–2648.

[42] ——, "Codes in the damerau distance for deletion and adjacent transposition correction," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2550–2570, 2018.

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2025.3619053

[43] K. Levick, R. Heckel, and I. Shomorony, "Achieving the capacity of a DNA storage channel with linear coding schemes," in *56th Annual Conference on Information Sciences and Systems (CISS)*, 2022, pp. 218–223.

[44] F. Palunvcić, D. Palunvcić, and B. T. Maharaj, "Capacity of Runlength-Limited and GC-content constrained codes for DNA data storage," in *IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 1937–1942.

[45] A. N. Ravi, A. Vahid, and I. Shomorony, "An information theory for out-of-order media with applications in DNA data storage," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 10, no. 2, pp. 334–348, 2024.

[46] A. Banerjee, Y. Yehezkeally, A. Wachter-Zeh, and E. Yaakobi, "Error-Correcting codes for Nanopore sequencing," *IEEE Transactions on Information Theory*, vol. 70, no. 7, pp. 4956–4967, 2024.

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

[48] H. Narayanan, P. Krishnan, and N. Parekh, "On achievable rates for the shotgun sequencing channel with erasures," in *IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 1730–1735.

[49] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA storage channels," in *IEEE Information Theory Workshop (ITW)*, 2015, pp. 1–5.

[50] ——, "Codes for DNA sequence profiles," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3125–3146, 2016.

[51] S. M. Hossein, T. Yazdi, H. M. Kiah, and O. Milenkovic, "Weakly mutually uncorrelated codes," in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 2649–2653.

[52] N. Beeri and M. Schwartz, "Improved rank-modulation codes for DNA storage with shotgun sequencing," *IEEE Transactions on Information Theory*, vol. 68, no. 6, pp. 3719–3730, 2022.

[53] O. Sabary, I. Preuss, R. Gabrys, Z. Yakhini, L. Anavy, and E. Yaakobi, "Error-correcting codes for combinatorial composite DNA," in *IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 109–114.

[54] F. Walter, O. Sabary, A. Wachter-Zeh, and E. Yaakobi, "Coding for composite DNA to correct substitutions, strand losses, and deletions," in *IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 97–102.

[55] K. A. Schouhamer Immink, K. Cai, T. T. Nguyen, and J. H. Weber, "Constructions and properties of efficient dna synthesis codes," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 10, no. 2, pp. 289–296, 2024.

[56] X. He, Y. Liu, T. Wang, and X. Tang, "Efficient explicit and pseudo-random constructions of constrained codes for DNA storage," *IEEE Transactions on Communications*, pp. 1–1, 2024.

[57] J. Liu, "Optimal RS codes and GRS codes against adversarial insertions and deletions and optimal constructions," *IEEE Transactions on Information Theory*, vol. 70, no. 9, pp. 6269–6279, 2024.

[58] F. Weindel, A. L. Gimpel, R. N. Grass, and R. Heckel, "Embracing errors is more effective than avoiding them through constrained coding for DNA data storage," in *59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2023, pp. 1–8.

[59] R. Nussinov and A. B. Jacobson, "Fast algorithm for predicting the secondary structure of single-stranded RNA," *of the National Academy of Sciences*, vol. 77, no. 11, pp. 6309–6313, 1980.

[60] P. Clote and R. Backofen, "Computational molecular biology: An introduction," in *Wiley Series in Mathematical and Computational Biology*, 2000.

[61] M. Zuker and D. Sankoff, "RNA secondary structures and their prediction," *Bulletin of Mathematical Biology*, vol. 46, no. 4, pp. 591–621, July 1984.

[62] M. K. Gupta, "On Some Linear Codes over $\mathbb{Z}_{2^s}$," Ph.D. dissertation, Indian Institute of Technology Kanpur, India, 1999.

[63] O. D. King, "Bounds for DNA codes with constant GC-content," *The Electronic Journal of Combinatorics*, vol. 10, article no. R33, 2003.

[64] K. G. Benerjee and A. Banerjee, "Bounds on size of homopolymer free codes," in *IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 809–814.

[65] T. T. Nguyen, K. Cai, H. M. Kiah, D. Tu Dao, and K. A. Schouhamer Immink, "On the design of codes for DNA computing: Secondary structure avoidance codes," in *IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 573–578.

[66] K. G. Benerjee and A. Banerjee, "On homopolymers and secondary structures avoiding, reversible, reversible-complement and GC-balanced DNA codes," in *IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 204–209.

[67] K. G. Benerjee and A. Banerjee, "On DNA codes with multiple constraints," *IEEE Communications Letters*, vol. 25, no. 2, pp. 365–368, 2021.

[68] Y. S. Kim and S. H. Kim, "New construction of DNA codes with constant-GC contents from binary sequences with ideal autocorrelation," in *IEEE International Symposium on Information Theory (ISIT)*, 2011, pp. 1569–1573.

APPENDIX

A. Computational Complexity Analysis for Algorithm 1

For a DNA sequence $\mathbf{z} = z_1 z_2 \dots z_{4t}$ of length $4t$, consider ℓ-length substrings $\mathbf{z}(i) = z_i z_{i+1} \dots z_{i+\ell-1}$, where $2 \leq \ell \leq \lfloor t/2 \rfloor$ and $1 \leq i \leq 4t - \ell + 1$. In Step 3 of Algorithm 1, two disjoint ℓ-length substrings $\mathbf{z}(i)$ and $\mathbf{z}(j)$ are compared for all $\mathbf{z} \in T^4$, with indices satisfying $1 \leq i \leq 4t - 2\ell + 1$, and $i+\ell \leq j \leq 4t-\ell+1$. For a fixed i, the number of valid j values is $(4t-\ell+1)-(i+\ell)+1 = 4t-2\ell-i+2$. Summing over all valid i, the total number of disjoint substring pairs for comparison is $\sum_{i=1}^{4t-2\ell+1}(4t-2\ell-i+2) = \frac{1}{2}(4t-2\ell+1)(4t-2\ell+2)$. Since comparisons are symmetric for SC/RCS conditions, the total number of substring comparisons per DNA string $\mathbf{z} \in T^4$ is twice this amount, i.e., $(4t-2\ell+1)(4t-2\ell+2)$. Next, for a fixed substring $z \in S$ (and hence $z \in T = \mathscr{A} \cup \{\mathbf{z}\}$) in Step 2, the total number of DNA strings in T^4 containing exactly k copies of z is $\binom{4}{k}(|T|-1)^{4-k}$, for $k = 1, 2, 3, 4$. Therefore, the total number of strings containing at least one z is $\sum_{k=1}^{4}\binom{4}{k}(|T|-1)^{4-k}$. Consequently, the total number of substring comparisons performed in Step 3 is $(4t-2\ell+1)(4t-2\ell+2)\sum_{k=1}^{4}\binom{4}{k}(|T|-1)^{4-k}$. For large $|T|$, this complexity is dominated by $\mathcal{O}\big((4t-2\ell)^2(|T|-1)^3\big)$. From Table I, the size of T is approximately bounded by $|T| \leq 6 \cdot 4^{t-2}$. Substituting this bound yields $(4t-2\ell+1)(4t-2\ell+2)\sum_{k=1}^{4}\binom{4}{k}\left(6\cdot 4^{t-2}-1\right)^{4-k}$. Accordingly, the overall complexity of substring comparisons in Algorithm 1 is $\mathcal{O}\big((4t-2\ell)^2\, 4^{3t-6}\big)$.