Cryptography and Communications (2021) 13:143–171
https://doi.org/10.1007/s12095-020-00459-7

On conflict free DNA codes

Krishna Gopal Benerjee[1] · Sourav Deb[1] · Manish K. Gupta[1]

Received: 28 October 2019 / Accepted: 24 September 2020 / Published online: 13 October 2020

Abstract

DNA storage has emerged as an important area of research. The reliability of a DNA storage system depends on designing those DNA strings (called DNA codes) that are sufficiently dissimilar. In this work, we introduce DNA codes that satisfy the newly introduced constraint, a generalization of the non-homopolymers constraint. In particular, each codeword of the DNA code has the specific property that any two consecutive sub-strings of the DNA codeword will not be the same. This is apart from the usual constraints such as Hamming, reverse, reverse-complement and GC-content. We believe that the new constraints proposed in this paper will provide significant achievements in reducing the errors, during reading and writing data into the synthetic DNA strings. We also present a construction (based on a variant of stochastic local search algorithm) to determine the size of the DNA codes with a constraint that each DNA codeword is free from secondary structures in addition to the usual constraint. This further improves the lower bounds from the existing literature, in some specific cases. A recursive isometric map between binary vectors and DNA strings is also proposed. By applying this map over the well known binary codes, we obtain classes of DNA codes with all of the above constraints, including the property that the constructed DNA codewords are free from the hairpin like secondary structures.

Keywords DNA codes · Homopolymers · Conflict free DNA strings · Hamming constraint · Reverse constraint · Reverse-complement constraint · GC-content constraint · Hairpin like secondary structures

Mathematics Subject Classification (2010) 68P30 · 94B60 · 94B65

The preliminary version of the paper is available at https://arxiv.org/abs/1902.04419

✉ Manish K. Gupta
mankg@computer.org

Krishna Gopal Benerjee
kg.benerjee@gmail.com

Sourav Deb
sourav_deb@daiict.ac.in

[1] Laboratory of Natural Information Processing, Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India

1 Introduction

The exponentially increasing demand in data storage forces to look into every possible option and DNA (DeoxyriboNucleic Acid) data storage has come out to be one of the most promising natural data storage for this purpose [11]. After the first striking implementation of large-scale archival DNA-based storage architecture by Church et al. [5] in 2012, followed by encoding scheme to DNA proposed by Goldman et al. [8] in 2013, researchers have taken great interests on the construction of DNA-based information storage systems [18, 20] because of its high storage density and longevity [5, 8, 36]. DNA consists of four types of bases or *nucleotides* (nts) called adenine (A), cytosine (C), guanine (G) and thymine (T) where, the Watson-Cricks complementary bases for A and C are T and G respectively and vice versa. In order to store data into synthetic DNA, we need to encode data into strings on quaternary alphabet $\{A, C, G, T\}$. The set of encoded DNA strings (also called as DNA codewords) on the quaternary alphabet is called DNA code. For a DNA string, the complement is a DNA string obtained by replacing each nucleotide by its complement. Similarly, for a DNA string, the reverse DNA string is a DNA string in the reverse order, and the complement of the reverse DNA string is called reverse-complement DNA string. The encoded strings are synthesized using DNA synthesizer for the purpose of writing into DNA strings and the synthesized DNA strings are stored in the appropriate environment. To extract the source data, the stored DNA strings are read using DNA sequencing.

Errors occur, particularly, during synthesis and sequencing the DNA strings, which can be reduced by choosing good encoding scheme for the DNA strings. Therefore, it is important to study the source of errors. Generally, insertion or deletion errors occur frequently in a DNA string with consecutive repetitions of a specific nucleotide (e.g. $AC\mathbf{GGGG}AT$) or of a block of nucleotides (e.g. $AG\mathbf{ATATAT}GC$) up to certain length [12, 21, 26, 32]. In addition, these DNA strings get misaligned more frequently, during DNA sequencing [26]. So we prefer DNA codes that exclude those codewords which contain consecutive repetition(s) of a specific nucleotide or a block of nucleotides. In this article, such DNA strings are defined as conflict free DNA strings. In literature, DNA codes without Homopolymers (DNA string with consecutive repetition of a nucleotide) [1, 2, 6, 10, 31] and without consecutive repeats of blocks [12, 16] are studied. As an extension, in this work, the considered conflict free DNA strings are not only free from Homopolymers but are also free from consecutive repetition of blocks of nucleotides.

In a single stranded DNA, the existence of two sub-strings where, the reverse-complement of one sub-string is the other one, results in forming an antiparallel double stranded hairpin like structure (also called hairpin loop or Stem-loop) by folding back upon itself [4, 13, 24, 27, 38].

An example of such hairpin like structure is illustrated in Fig. 1. For DNA sequencing, it is preferred to avoid such secondary structures [13]. In this work, all the codewords of the constructed conflict free DNA codes are free from hairpin like structures with stem length more than 2.

A DNA string can be read using specific hybridization between the DNA string and its complement DNA string [23]. If the DNA strings in a code are not different enough among themselves then nonspecific hybridization will occur and it will be a prominent cause of error. Therefore, a set of DNA codewords is preferred in which DNA strings are sufficiently different among themselves. Hamming distance between two strings of same length over the same alphabets is the number of positions in which the symbols in the strings are different. So, construction of DNA code with Hamming constraint (ensures the difference among

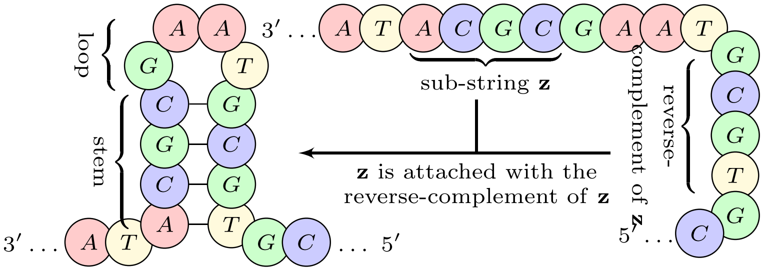

Fig. 1 An example of hairpin like secondary structures in a single stranded DNA

DNA codewords), reverse constraint (ensures the difference between DNA codewords and their reverse DNA strings), and reverse-complement constraint (ensures the difference between DNA codewords and their reverse-complement DNA strings) is preferred. In literature, DNA codes with reverse and reverse-complement constraints are constructed from finite fields and finite rings in [9, 17, 29, 33].

The thermal stability of a DNA string depends on the GC-content (the total number of $G's$ and $C's$) in the DNA string [35]. On the other hand, the high GC-content leads to the insertion and deletion error during polymerase chain reaction (PCR). Therefore, such DNA codes are preferred in which each DNA codeword has the same GC-content and equal to almost half of its length and the constraint for the DNA codes is called GC-content constraint. In [14, 15, 30, 33], DNA codes with balanced GC content are studied. DNA codes with reverse, reverse-complement and GC-content constraints are studied in [7, 30]. In [3], a revised lower bound on size of DNA codes with GC-content and reverse-complement constraints are obtained. In fact, DNA codes with balanced GC content and without Homopolymers are also studied in [6, 10, 31].

In this paper, we have studied the DNA codes with multiple properties such as each DNA string is free from consecutive reparation of DNA blocks up to a certain length. Any DNA codeword of the DNA code sufficiently differs from other DNA codewords, reverse of DNA codewords and reverse-complement DNA codewords. In addition, each codeword of the DNA code is also free from hairpin like secondary structures. These properties significantly help to reduce bit-flip, insertion, and deletion errors simultaneously during reading and writing DNA strings. The lower bounds on the maximum size of DNA code with all those properties are also obtained in this paper. To the best of author's knowledge, in literature, an algebraic solution for DNA codes with all the constraints is not studied yet. In this work, an algebraic structure for the family of DNA codes is proposed where the constructed DNA codes meet all the constraints such as Hamming, reverse, reverse-complement, and GC-content constraints. Apart from that, all the DNA codewords do not have any consecutive identical sub-string(s) up to a certain length ℓ (generalization of non-homopolymer constraint [12, 21, 26, 32]. These are known as non-homopolymer constraint of order ℓ). In addition, these codewords are free from hairpin like secondary structures. In this paper, an algorithm is given which calculates the DNA code with the property that each DNA codeword does not have any consecutive repeated sub-string of any length. In addition, DNA codes with Hamming constraint, reverse constraint, reverse-complement constraint, and GC-content constraint are obtained. For a DNA code with all the constraints, the obtained code size is improved for some specific parameters as given in [19, Table I]. Further, family of DNA codes have been obtained with Hamming, reverse, reverse-complement, and

GC-content constraints where, each DNA codeword is free from hairpin like secondary structure and non-homopolymer constraint.

In Section 2, preliminary for DNA codes is discussed. Complete conflict free DNA codes with all the constraints are studied in Section 3. A recursive mapping from binary strings to DNA strings is discussed in Section 4, which is also an isometry between non-homopolymer distance over binary strings and Hamming distance over DNA strings. The conditions on binary strings are obtained, which ensure the constraints on encoded DNA strings in the same section. In Section 5, a family of DNA codes are obtained from binary Reed-Muller codes. Section 6 concludes the work with general remarks.

2 Preliminary

A code $\mathcal{C}$ with parameters (n, M, d) over an alphabet Σ of size q is a set of M distinct strings (also called codewords) with length n such that the distance between any two distinct strings is at least d. Codes over $\{0, 1\}$ are called binary codes. Similarly, codes over an alphabet of size 4 are known as Quaternary codes. In particular, codes over alphabet $\Sigma_{DNA} = \{A, C, G, T\}$ are called DNA codes (denoted by $\mathcal{C}_{DNA}$) respectively. For various applications, codes with various distances (such as Gau distance [17]) are studied in literature. In this work, DNA codes with Hamming distance and binary codes with a newly defined distance are studied. For any strings $\mathbf{x}$ and $\mathbf{y}$ in Σ^n, the Hamming distance $d_H(\mathbf{x}, \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$ is the total number of positions at which they differ. For a code $\mathcal{C} \subset \Sigma^n$, the minimum Hamming distance is $d_H = \min\{d_H(\mathbf{x}, \mathbf{y}) : \mathbf{x} \neq \mathbf{y} \text{ and } \mathbf{x}, \mathbf{y} \in \mathcal{C}\}$. Note that one can find fields or rings over the alphabet Σ. For a ring defined over the alphabet Σ, a code on Σ is called linear if the code is sub-module over the ring. For a field defined over the alphabet Σ, a code on Σ is called linear if the code is row span of a matrix over the field. For any linear code, the matrix is called the generator matrix if rows of that matrix are linearly independent.

For any DNA string $\mathbf{x} = (x_1\ x_2 \ldots x_n) \in \Sigma_{DNA}^n$, the reverse, complement and reverse-complement DNA strings of $\mathbf{x}$ are $\mathbf{x}^r = (x_n\ x_{n-1} \ldots x_1)$, $\mathbf{x}^c = (x_1^c\ x_2^c \ldots x_n^c)$, and $\mathbf{x}^{rc} = (x_n^c\ x_{n-1}^c \ldots x_1^c)$ respectively where, $A^c = T$, $C^c = G$, $G^c = C$ and $T^c = A$. As defined in [23], for any DNA code $\mathcal{C}_{DNA}$ with parameter (n, M, d_H), the various constraints are defined as follows:

- *Hamming constraint*: The Hamming distance $d_H(\mathbf{x}, \mathbf{y}) \geq d_H$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}_{DNA}$ and $\mathbf{x} \neq \mathbf{y}$.
- *Reverse constraint*: The Hamming distance $d_H(\mathbf{x}, \mathbf{y}^r) \geq d_H$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}_{DNA}$ and $\mathbf{x} \neq \mathbf{y}^r$.
- *Reverse-complement constraint*: The Hamming distance $d_H(\mathbf{x}, \mathbf{y}^{rc}) \geq d_H$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}_{DNA}$ and $\mathbf{x} \neq \mathbf{y}^{rc}$.
- *GC-content constraint*: If the total number of G's and C's in each codeword is same and equal to g then the code satisfies g-GC-content constraint. For a specific case $g = \lfloor n/2 \rfloor$, the $\lfloor n/2 \rfloor$-GC-content constraint is simply called GC-content constraint.

Consider a DNA code $\mathcal{C}_{DNA}$ with the minimum Hamming distance d_H. For each $\mathbf{x} \in \mathcal{C}_{DNA}$, if $\mathbf{x}^r \in \mathcal{C}_{DNA}$ then, from the distance property of code, $d_H(\mathbf{y}, \mathbf{x}^r) \geq d_H$ for each $\mathbf{y} \in \mathcal{C}_{DNA}$ such that $\mathbf{x}^r \neq \mathbf{y}$. Therefore, the code satisfies the reverse constraint. Similarly, for each $\mathbf{x} \in \mathcal{C}_{DNA}$, if $\mathbf{x}^{rc} \in \mathcal{C}_{DNA}$ then from the distance property of code again, $d_H(\mathbf{y}, \mathbf{x}^{rc}) \geq d_H$ for each $\mathbf{y}$ $(\neq \mathbf{x}^{rc})$ in $\mathcal{C}_{DNA}$, and hence the code satisfies the reverse-complement constraint. In consequence, researchers are curious in the construction of DNA codes which are closed under reverse and reverse-complement DNA strings [17]. Thus, motivated by this, we construct a set of DNA strings for a given length such that those DNA strings satisfy multiple constraints. In the following lemma, the distinct DNA strings of fix length with some additional constraints are enumerated.

Lemma 1 *For a given length n,*

1. *there exists $4^{\lceil n/2\rceil}$ number of distinct DNA strings $\boldsymbol{x} \in \Sigma_{DNA}^n$ such that $\boldsymbol{x} = \boldsymbol{x}^r$,*
2. *for an even n, there exists $4^{n/2}$ number of DNA strings $\boldsymbol{x} \in \Sigma_{DNA}^n$ such that $\boldsymbol{x} = \boldsymbol{x}^{rc}$,*
3. *for a positive integer m ($\leq n$), there exists $\binom{n}{m}2^n$ distinct DNA strings $\boldsymbol{x} \in \Sigma_{DNA}^n$ each with GC-content m,*
4. *for an even positive integer m ($\leq n$), there exists $\binom{n/2}{m/2}2^{n/2}$ distinct DNA strings $\boldsymbol{x} \in \Sigma_{DNA}^n$ each with GC-content m and $\boldsymbol{x} = \boldsymbol{x}^{rc}$ where, n is even, and*
5. *for a positive integer m ($\leq n$), there exists η distinct DNA strings $\boldsymbol{x} \in \Sigma_{DNA}^n$ each with m - GC-content and $\boldsymbol{x} = \boldsymbol{x}^r$ where,*

$$\eta = \begin{cases} 0 & \textit{if n is even and m is odd,} \\ \binom{\lfloor n/2\rfloor}{\lfloor m/2\rfloor}2^{\lceil n/2\rceil} & \textit{otherwise.} \end{cases}$$

Proof Consider $\mathbf{x} = (x_1\ x_2 \ldots x_n) \in \Sigma_{DNA}^n$.

1) If $\mathbf{x} = \mathbf{x}^r$ then, $x_i = x_{n-i+1}$ for $i = 1, 2, \ldots, \lceil n/2\rceil$. Therefore, there exists $4^{\lceil n/2\rceil}$ number of distinct DNA strings $\mathbf{x} \in \Sigma_{DNA}^n$ such that $\mathbf{x} = \mathbf{x}^r$.
2) If $\mathbf{x} = \mathbf{x}^{rc}$ then, $x_i = x_{n-i+1}^c$ for $i = 1, 2, \ldots, \lfloor n/2\rfloor$. Therefore, there exists $4^{n/2}$ number of distinct DNA strings $\mathbf{x} \in \Sigma_{DNA}^n$ such that $\mathbf{x} = \mathbf{x}^{rc}$. Note that, for any positive odd n, any $\mathbf{x} \in \Sigma_{DNA}^n$ can not equal $\mathbf{x}^{rc}$.
3) There are $\binom{n}{m}2^m$ ways to fill m positions out of n positions by a symbol from $\{C, G\}$. If the remaining n - m positions are filled by a symbol from $\{A, T\}$ then there are $\binom{n}{m}2^n$ distinct DNA strings each with GC-content m.

Proofs of remaining results are similar. □

For theoretical analysis, we define complement constraint for a DNA code similar to reverse and reverse-complement constraints. A DNA code $\mathcal{C}_{DNA}$ satisfies the complement constraint, if, for any $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{C}_{DNA}$ such that $\mathbf{x} \neq \mathbf{y}^c$, $d_H(\mathbf{x}, \mathbf{y}^c) \geq d_H$.

Remark 1 Consider a DNA code $\mathcal{C}_{DNA}$ with the minimum Hamming distance d_H. For any codeword $\mathbf{x} \in \mathcal{C}_{DNA}$, if $\mathbf{x}^c$ and $\mathbf{x}^r$ are also in the DNA code $\mathcal{C}_{DNA}$ then, from the distance property of code, the DNA code satisfies reverse, complement, and reverse-complement constraints.

Remark 2 Consider a DNA code $\mathcal{C}_{DNA}$ with the minimum Hamming distance d_H such that $\mathbf{x}^c \in \mathcal{C}_{DNA}$ for each codeword $\mathbf{x} \in \mathcal{C}_{DNA}$. For any codeword $\mathbf{x}, \mathbf{y} \in \mathcal{C}_{DNA}$, if $d_H(\mathbf{x}, \mathbf{y}^r) \geq d_H$ then, from the reverse and reverse-complement properties of DNA, the DNA code satisfies reverse, complement, and reverse-complement constraints.

Remark 3 Consider a DNA code $\mathcal{C}_{DNA}$ with the minimum Hamming distance d_H such that $\mathbf{x}^r \in \mathcal{C}_{DNA}$ for each codeword $\mathbf{x} \in \mathcal{C}_{DNA}$. For any codeword $\mathbf{x}, \mathbf{y} \in \mathcal{C}_{DNA}$, if $d_H(\mathbf{x}, \mathbf{y}^c) \geq d_H$ then, from the reverse and reverse-complement properties of DNA, the DNA code satisfies reverse, complement, and reverse-complement constraints.

For a DNA code with some properties, the following lemma ensures the reverse-complement constraint.

Lemma 2 *For a DNA code of parameter (n, M, d_H) with reverse and complement constraints, if $d_H \leq n/2$ then the DNA code will satisfy the reverse-complement constraint.*

Proof For any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $d_H(\mathbf{y}^{rc}, \mathbf{x}^c) = d_H(\mathbf{y}^r, \mathbf{x})$. So, $d_H(\mathbf{x}, \mathbf{x}^c) \leq d_H(\mathbf{x}, \mathbf{y}^{rc}) + d_H(\mathbf{y}^{rc}, \mathbf{x}^c) = d_H(\mathbf{x}, \mathbf{y}^{rc}) + d_H(\mathbf{y}^r, \mathbf{x})$. This implies, $d_H(\mathbf{x}, \mathbf{x}^c) \leq d_H(\mathbf{x}, \mathbf{y}^{rc}) + d_H$. But, $d_H(\mathbf{x}, \mathbf{x}^c) = n$, so, $n - d_H \leq d_H(\mathbf{x}, \mathbf{y}^{rc})$. Therefore, if $d_H \leq n/2$ then $d_H(\mathbf{x}, \mathbf{y}^{rc}) \geq d_H$. □

In a single stranded DNA, if there exist two sub-strings such that one sub-string is the reverse-complement of another sub-string then the DNA strand folds back and attaches the both sub-strings to each other and forms hairpin like secondary structures with stems and loops of certain length [4, 13, 24, 27]. The stem size of more than 2 bases long reasonably approximates the hairpin like structures. The single stranded DNA AT**ACGC**$GAAT$**GCGT**GC, considered in Fig. 1, contains the reverse-complementary sub-strings $ACGC$ and $TGCG$ (see the bold sub-strings). The sub-strings are attached to each other and forms a stem of length 4 base pairs, and a loop of size 4 bases long. Therefore, in this work, DNA strings are considered to be free from hairpin like structures with stem length of more than 2 bases long.

Definition 1 A DNA string is called free from secondary structures (of stem length more than 2) if the DNA string does not contain any two sub-strings of length more than 2 such that one is the reverse-complement of the other.

Remark 4 Consider a DNA string of length n (> 5) which contains two sub-strings of length t $(n \geq t \geq 4)$ such that one sub-string is the reverse-complement of the other. Then, the DNA string will also contain two sub-strings of length 3 such that one is the reverse-complement of the other. Therefore, it is sufficient for a DNA string, free from secondary structures of stem length more than 2, not to contain two sub-strings which are reverse-complement to each other.

Apart from the pairing between Watson-Crick complement base pairs, the secondary structures can also be formed with Wobble base pairs or cross hybridization between base pairs in a single strand DNA [4, 13, 24, 27, 38]. In this work, we have considered only those secondary structures which have stem length more than 2 and each bounded base pair in the stem is Watson-Crick complement pairs. We called such DNA strings free from secondary structures in this paper.

Remark 5 A DNA string is free from secondary structures if and only if the complement of the DNA string is also free from secondary structures. The same result holds for reverse and reverse-complement.

For example, the DNA string $ACATCG$ is free from reverse-complement sub-strings of length 3 because the reverse-complement of ACA is TGT and TGT is not a sub-string of $ACATCG$. The reverse and reverse-complement DNA strings $GCTACA$ and $CGATGT$ are also free from reverse-complement sub-strings. Therefore, the following remark holds.

Remark 6 Any DNA string is free from secondary structures of stem length more than 2 if and only if it is free from secondary structures of stem length 3.

3 On complete conflict free DNA strings

It is evident that, the process of sequencing and synthesizing of DNA strings will be erroneous due to the existence of homopolymers and consecutive repetition of same sub-string(s) of certain length in DNA [12, 21, 26, 32]. So it is preferable to construct DNA codes in such a way that each DNA codeword will be free from Homopolymers or consecutive repetition(s) of same sub-string(s) of certain length. In general, a sequence which is free from the consecutive repetition of a block is known as square free sequence or repeat free sequence in literature [22]. Thus motivated by this, we define ℓ conflict free DNA strings and ℓ conflict free DNA code. Also the necessary and sufficient condition for a DNA string to be ℓ conflict free is given.

Definition 2 For positive integers n and ℓ $(\leq n/2)$, a DNA string is called ℓ conflict free, if the DNA string of length n is free from consecutive repetition(s) of identical sub-string(s) of length t for each $t = 1, 2, \ldots, \ell$.

For example, the DNA string $ATCATCG$ is 2 conflict free but not 3 conflict free as the substring ATC has a consecutive repetition in the DNA string while any two consecutive sub-strings of same length (≤ 2) are not same, *i.e.*, the DNA string does not contain any of the DNA sub-strings $ATAT$, $TCTC$, $CACA$, $CGCG$, AA, TT, CC and GG. Note that, 1 conflict free DNA strings are also known as DNA strings free from homopolymers in literature [1, 2, 6, 10, 31]. Also note, for any positive integer ℓ (≥ 2), an ℓ conflict free DNA string is also $\ell - 1$ conflict free.

Remark 7 A DNA string is ℓ conflict free if and only if the complement DNA string is also ℓ conflict free. Note that the result also holds for reverse and reverse-complement DNA strings.

Definition 3 For positive integers n and ℓ $(\leq \lfloor n/2 \rfloor)$, a DNA code with length n is called ℓ conflict free DNA code if each DNA codeword of the DNA code is ℓ conflict free.

For example, the DNA code $\{ACTG, TGAC, CAGT, GTCA\}$ is 2 conflict free DNA code, since each DNA codeword is 2 conflict free DNA string.

One can observe that the maximum length of a DNA sub-string will be $\lfloor n/2 \rfloor$, which can be repeated in a DNA string of length n. Hence, from Definition 2, a DNA string will be free from repetition(s) of DNA sub-string(s) of any length, if it is $\lfloor n/2 \rfloor$ conflict free. For a positive integer n, let $S(n)$ be the set of all $\lfloor n/2 \rfloor$ conflict free DNA strings each of length n. For any $\mathbf{z} \in S(n)$ we will have $\mathbf{z}^r, \mathbf{z}^c, \mathbf{z}^{rc} \in S(n)$. Also note that, any DNA code $\mathcal{C}_{DNA} \subset S(n)$ is always a $\lfloor n/2 \rfloor$ conflict free DNA code.

Various computational approaches to construct DNA codes with some additional constraints are studied in literature [3, 28, 33, 34, 37]. In this work, ℓ conflict free DNA codes are constructed using stochastic local search in a seed set of ℓ conflict free DNA strings such that each DNA string has a fix GC-content and all the DNA strings are free from reverse-complement sub-strings. The computational construction for ℓ conflict free DNA codes is given as follows.

Construction 1 For given positive integers n, ℓ ($\leq \lfloor n/2 \rfloor$) and g ($\leq n$), let $\mathcal{S} \subset \Sigma_{DNA}^{n}$ be the set of all ℓ conflict free DNA strings such that GC-content of each DNA string is g and every DNA string is free from secondary structures. For a sub-set R of random cardinality and containing DNA strings which are randomly selected from $\mathcal{S}$, the DNA code $\mathcal{C}_{DNA} = R \cup \{\mathbf{x}^{r}, \mathbf{x}^{c}, \mathbf{x}^{rc} : \mathbf{x} \in R\}$ is an ℓ conflict free DNA code with Hamming, reverse and reverse-complement constraints where, each DNA codeword is free from reverse-complement substrings and GC-constant of each DNA codeword is fix g.

For example, consider $n = 3$, $\ell = 1$ and $g = 2$. The seed set will be $\mathcal{S} = \{ACG, AGC, CAC, CAG, CGA, CGT, CTC, CTG, GAC, GAG, GCA, GCT, GTC, GTG, TCG, TGC\}$. Note that, for $R = \{CAC, CGT, ACG, TGC\}$, the 1 conflict free DNA code with GC-content, reverse and reverse-complement constraints is $\mathcal{C}_{DNA} = \{CAC, CGT, ACG, TGC, GCA, GTG\}$ where, each DNA codeword is free from secondary structures. Also note, the DNA code size and the minimum Hamming distance of the code are $M = 6$ and $d_H = 2$.

For a given DNA string, the computational complexity to determine if the DNA string is ℓ conflict free (using Definition 2) is more than the computational complexity to determine whether the DNA string is free from secondary structures (using Remark 6), and, it is again more than the computational complexity to determine the GC-content of the DNA string. Therefore, in order to construct the seed set $\mathcal{S}$ for the Construction 1, one can reduce the computations by removing DNA strings in the following order.

1. Remove all the DNA strings from the complete set Σ_{DNA}^{n} which do not have g-GC-content.
2. Remove all the DNA strings from the remaining set which are not free from secondary structures of stem length 3.
3. Remove all the DNA strings from the remaining set which are not ℓ conflict free.

For a given length n and Hamming distance d_H, the maximum size of code is subject to interest among researches. Now, similar to [23], some notations for the maximum size of ℓ conflict free DNA codes are introduced here.

– For a DNA code of codeword length n and the minimum distance d_H with GC-content and reverse-complement constraints, the maximum size of the DNA code is denoted by $A_4^{GC,rc}(n, d_H)$ where, the terms GC and rc stand for GC-content and reverse-complement constraints.
– The maximum size of an ℓ conflict free DNA code with length n and the minimum Hamming distance d_H is $A_4^{cf}(n, d_H, \ell)$. So, $A_4^{cf,GC}(n, d_H, 1)$ denotes the maximum size of a DNA code of codeword length n and the minimum distance d_H with GC-content constraint such that each codeword is free from homopolymers where, the term cf stands for conflict free property of the DNA code.
– For an ℓ conflict free DNA code of codeword length n and the minimum distance d_H with reverse constraint such that each DNA codeword is free from secondary structures (hairpin like structure) of length more than 2, the maximum size of the DNA code is denoted by $A_4^{cf,hf,r}(n, d_H, \ell)$ where, the terms hf and r stand for property free from secondary structures and reverse constraint.
– For an ℓ conflict free DNA code of codeword length n and the minimum distance d_H with reverse-complement constraint, $A_4^{cf,hf,rc}(n, d_H, \ell)$ denotes the maximum size of

the DNA code where, each DNA codeword is free from secondary structures of length more than 2.

- For any ℓ conflict free DNA codes of codeword length n and the minimum distance d_H with GC-contant constraint, the maximum size of the DNA code is denoted by $A_4^{cf,hf,GC}(n, d_H, \ell)$ where, each DNA codeword is free from secondary structures of length more than 2.
- For any ℓ conflict free DNA codes of codeword length n and the minimum distance d_H with GC-contant, reverse, and reverse-complement constraints, the maximum size of the DNA code is denoted by $A_4^{cf,hf,GC,r,rc}(n, d_H, \ell)$ where, each DNA codeword is free from secondary structures of length more than 2.

The relations among sizes of ℓ conflict free DNA codes with additional constraints are given in following theorem.

Theorem 1 *For a positive even integer n,*

$$A_4^{cf,hf,GC,r}(n, d_H, \ell) = A_4^{cf,hf,GC,rc}(n, d_H, \ell),$$

and for a positive odd integer n,

$$\begin{aligned} A_4^{cf,hf,GC,r}(n, d_H + 1, \ell) &\leq A_4^{cf,hf,GC,rc}(n, d_H, \ell) \\ &\leq A_4^{cf,hf,GC,r}(n, d_H - 1, \ell). \end{aligned}$$

Proof Similar to the proof of [23, Theorem 4.1], consider an ℓ conflict free DNA code $\mathcal{C}_{DNA}$ with codeword length n and the minimum Hamming distance d_H. In addition, each codeword of the DNA code is free from secondary structures and has fixed GC-content. Now, for even n, the DNA code $\mathcal{C}_{DNA} = \{\mathbf{a}_i\mathbf{b}_i\}$ satisfies reverse constraint where, the sizes of $\mathbf{a}_i$ and $\mathbf{b}_i$ are same and $\mathbf{a}_i\mathbf{b}_i$ is the concatenation of sequences $\mathbf{a}_i$ and $\mathbf{b}_i$ in same order. Then the DNA code $\mathcal{C}'_{DNA} = \{\mathbf{a}_i\mathbf{b}_i^c\}$ satisfies reverse-complement constraint. Observe that the parameters of both the DNA codes $\mathcal{C}_{DNA}$ and $\mathcal{C}'_{DNA}$ are same and therefore, $A_4^{cf,hf,GC,r}(n, d_H, \ell) \leq A_4^{cf,hf,GC,rc}(n, d_H, \ell)$. Similarly, one can prove the inequality $A_4^{cf,hf,GC,r}(n, d_H, \ell) \geq A_4^{cf,hf,GC,rc}(n, d_H, \ell)$. Thus, the result follows for even n. Again, for odd n, consider a ℓ conflict free DNA code $\mathcal{C}_{DNA}$ with parameter $(n, A_4^{cf,hf,GC,r}(n, d_H + 1, \ell), d_H + 1)$ with reverse and GC-content constraints. Also each codeword of the DNA code is free from secondary structures. For some $x \in \Sigma_{DNA}$, $\mathcal{C}_{DNA} = \{\mathbf{a}_i x \mathbf{b}_i\}$ where, the sizes of $\mathbf{a}_i$ and $\mathbf{b}_i$ are same and $\mathbf{a}_i\mathbf{b}_i$ is the concatenation of sequences $\mathbf{a}_i$ and $\mathbf{b}_i$ in same order. Now, consider $\mathcal{C}^*_{DNA} = \{\mathbf{a}_i\mathbf{b}_i\}$ of even length which is obtained by deleting the middle symbol x of each codeword of the DNA code $\mathcal{C}_{DNA}$. Therefore, from even case, $A_4^{cf,hf,GC,r}(n, d_H + 1, \ell) \leq A_4^{cf,hf,GC,rc}(n - 1, d_H, \ell)$. The first inequality for odd n follows from $A_4^{cf,hf,GC,rc}(n - 1, d_H, \ell) \leq A_4^{cf,hf,GC,rc}(n, d_H, \ell)$. Similarly, one can prove the second inequality for odd n. □

Remark 8 Consider an ℓ conflict free DNA code with codeword length n and the minimum Hamming distance d_H.

- For positive integers n, ℓ $(< \lfloor n/2 \rfloor)$ and d_H $(\leq n)$,

$$\begin{aligned} A_4^{cf}(n, d_H, \ell) &\geq A_4^{cf}(n, d_H, \ell + 1) \\ A_4^{cf,hf}(n, d_H, \ell) &\geq A_4^{cf,hf}(n, d_H, \ell + 1) \text{ and} \\ A_4^{cf}(n, d_H, \ell) &\geq A_4^{cf,hf}(n, d_H, \ell) \\ &\geq A_4^{cf,hf,GC}(n, d_H, \ell) \\ &\geq A_4^{cf,hf,GC,r}(n, d_H, \ell) \\ &\geq A_4^{cf,hf,GC,r,rc}(n, d_H, \ell). \end{aligned}$$

- For positive integers n and d_H $(\leq n)$,

$$\begin{aligned} &A_4^{cf,GC}(n, d_H, 1) \geq A_4^{cf,hf,GC}(n, d_H, \ell), \text{ and} \\ &A_4^{GC,rc}(n, d_H) \geq A_4^{cf,hf,GC,rc}(n, d_H, \ell). \end{aligned}$$

For $n = 1, 2, \ldots, 10$, $g = \lfloor n/2 \rfloor$ and $\ell = 1, 2, \ldots, \lfloor n/2 \rfloor$, the DNA codes with various parameters are calculated using Construction 1 and the maximum value of obtained code sizes are listed in Table 1 where, the DNA codes satisfy reverse, reverse-complement, GC-content constraints, and each codeword of the DNA code is $\lfloor n/2 \rfloor$ conflict free and also free from secondary structures. For given n and ℓ, first, the seed set $\mathcal{S}$ is obtained such that each DNA string in the set contains $\lfloor n/2 \rfloor$ GC-content. Further, for a random subset R of the set $\mathcal{S}$, the $\lfloor n/2 \rfloor$ conflict free DNA code $\mathcal{C}_{DNA} = R \cup \{\mathbf{x}^r, \mathbf{x}^c, \mathbf{x}^{rc} : \mathbf{x} \in \mathcal{C}\}$ is obtained. From the seed set $\mathcal{S}$, the sub-set R is generated 10^6 times and, for each subset R, the size $|\mathcal{C}_{DNA}|$ is enumerated. For given n and d_H, the maximum value among all the enumerated $|\mathcal{C}_{DNA}|$ (lower bound of $A_4^{cf,hf,GC,r,rc}(n, d_H, \lfloor n/2 \rfloor)$) is listed in Table 1. Note that $A_4^{cf,hf,GC}(n, 1, \lfloor n/2 \rfloor) = A_4^{cf,hf,GC,r,rc}(n, 1, \lfloor n/2 \rfloor)$ and $A_4^{cf,hf}(n, 1, \lfloor n/2 \rfloor) = A_4^{cf,hf,r,rc}(n, 1, \lfloor n/2 \rfloor)$, for a positive integer n. Therefore, from Remark 7 and Definition 3,

Table 1 Lower bound of $A_4^{cf,hf,GC,r,rc}(n, d_H, \lfloor n/2 \rfloor)$ (the maximum size of $\lfloor n/2 \rfloor$ conflict free DNA code with Hamming, reverse, reverse-complement and GC-content constraints, and each DNA codeword is free from secondary structures) are given. Values given in circles are improved from the values given in [19, Table I] or [3, Table II]. The bold values are those values which archive the values given in [19, Table I] or [3, Table II]

n \ d_H	1	2	3	4	5	6	7	8	9	10
2	8	4	–	–	–	–	–	–	–	–
3	16	6	2	–	–	–	–	–	–	–
4	48	**32**	**⑫**	**④**	–	–	– –	–	–	–
5	108	48	14	**④**	**②**	–	–	–	–	–
6	320	92	32	**⑳**	0	**④**	–	–	–	–
7	656	136	52	28	10	**④**	**②**	–	–	–
8	1832	220	88	48	20	**⑫**	0	**④**	–	–
9	3640	306	124	64	32	**⑯**	6	**④**	**②**	–
10	9504	564	200	104	52	28	**⑯**	**⑧**	0	**④**

$A_4^{cf,GC,r,rc}(n, 1, \lfloor n/2 \rfloor) = |\mathcal{S}|$ where, $|\mathcal{S}|$ is cardinality of seed set in Construction 1. Therefore, for $n = 2, 3, \ldots, 10$, the listed lower bounds of $A_4^{cf,hf,GC,r,rc}(n, 1, \lfloor n/2 \rfloor)$ are tight. Similarly, from the definition of Hamming distance one can observe that,

$$A_4^{cf,hf,GC,r,rc}(n, n, \lfloor n/2 \rfloor) = \begin{cases} 2 & \text{if } n \text{ is odd,} \\ 4 & \text{if } n \text{ is even.} \end{cases}$$

Recall, for any odd length DNA string $\mathbf{x}$, $d_H(\mathbf{x}, \mathbf{x}^r) < n$, and hence, the values $A_4^{cf,hf,GC,r,rc}(n, n, \lfloor n/2 \rfloor)$ are also tight for $d_H = n$ (= 2, 3 . . . , 10) in Table 1. Values which are written in bold font in the table indicate equal or improved values from [19, Table I] or [3, Table II], and values written in a circle in the table are improved values from [19, Table I] or [3, Table II] . For the equal and improved values, the corresponding code parameters have been considered and the respective DNA codes and their codewords are listed in Table 6. Apart from the existing literature, all the specified constraints have been considered in Table 1. Some of those values are better when compared with the DNA code size with less constraints listed in Table 2.

Note that the values in Table I, [19] are the lower bounds for the maximum size of DNA codes satisfying GC-content constraint and free from Homopolymers, and in Table II, [3] are lower bounds for the maximum size of DNA codes satisfying GC-content and reverse-complement constraints, on the other hand, values listed in Table 1 in this paper are the lower bounds for the maximum size of complete conflict free DNA codes satisfying Hamming, reverse, reverse-complement and GC-content constraints. From Remark 8 and Table 2, some new lower bounds are obtained for $A_4^{cf,GC}(n, d_H, 1)$ and $A_4^{GC,rc}(n, d_H)$. For $A_4^{cf,GC}(n, d_H, 1)$, the newly updated bounds are

$$A_4^{cf,GC}(4, 3, 1) \geq 12,$$
$$A_4^{cf,GC}(6, 4, 1) \geq 20,$$
$$A_4^{cf,GC}(8, 6, 1) \geq 12,$$
$$A_4^{cf,GC}(9, 6, 1) \geq 16,$$
$$A_4^{cf,GC}(9, 9, 1) \geq 2,$$
$$A_4^{cf,GC}(10, 7, 1) \geq 16, \text{ and}$$

Table 2 List of improved lower bound of $A_4^{cf,hf,GC,r,rc}(n, d_H, \lfloor n/2 \rfloor)$ (the maximum size of $\lfloor n/2 \rfloor$ conflict free DNA code with Hamming, reverse, reverse-complement and GC-content constraints and each DNA string is free from secondary structures) from [19, Table I] and [3, table II]

DNA code Parameters (n, d_H)	Lower bound of $A_4^{cf,hf,GC,r,rc}(n, d_H, \lfloor n/2 \rfloor)$ Table 1	Lower bound of $A_4^{cf,GC}(n, d_H, 1)$ Table I in [19]	Lower bound of $A_4^{GC,rc}(n, d_H)$ Table II in [3]
(4, 3)	12	11	11
(6, 4)	20	16	16
(8, 6)	12	9	12
(9, 6)	16	15	20
(9, 9)	2	0	1
(10, 7)	16	7	16
(10, 8)	8	5	8

Table 3 Non-Homopolymer Map and an Example with $\mathbf{x} = CG$ and $\mathbf{y} = AT$

Binary	Previous Nucleotide Block			
Digit	$\mathbf{x}$	$\mathbf{x}^c$	$\mathbf{y}$	$\mathbf{y}^c$
0	$\mathbf{y}$	$\mathbf{y}^c$	$\mathbf{x}^c$	$\mathbf{x}$
1	$\mathbf{y}^c$	$\mathbf{y}$	$\mathbf{x}$	$\mathbf{x}^c$

(a) Non-Homopolymer Map

Binary	Previous Nucleotide Block			
Digit	CG	GC	AT	TA
0	AT	TA	GC	CG
1	TA	AT	CG	GC

(b) An example of Non-Homopolymer map

$$A_4^{cf,GC}(10, 8, 1) \geq 8.$$

Similarly, for $A_4^{GC,rc}(n, d_H)$, the newly achieved bounds are

$$A_4^{GC,rc}(4, 3) \geq 12,$$
$$A_4^{GC,rc}(6, 4) \geq 20, \text{ and}$$
$$A_4^{GC,rc}(9, 9) \geq 2.$$

4 Mappings and their properties

For a positive integer ℓ, the frequency of occurrence of insertion or deletion errors in an $\ell + 1$ conflict free DNA string is less in an ℓ conflict free DNA string. Therefore, the chances of occurrence of these errors is significantly low in an ℓ conflict free DNA string for a sufficiently large ℓ. On the other hand, the computational complexity of Construction 1 is high. Therefore, to sidestep the computational approach, a recursive mapping is defined algebraically in this section which ensures that the obtained DNA strings will be ℓ conflict free. Moreover, DNA codes satisfying all the constraints are also studied with respect to the mapping in this section.

Definition 4 (Non-Homopolymer map of order ℓ): For a positive integer ℓ, consider $\mathbf{x}, \mathbf{y} \in \Sigma_{DNA}^{\ell}$ such that $\mathbf{x} \neq \mathbf{y}$. Define a map $f : \{0, 1\} \times \{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\} \rightarrow \{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\}$ such that Table 3(a) holds.

Table 4 Parameters for conflict free DNA codes encoded from binary codes

Binary code	DNA code parameters		
(Encoded from)	$n\ell$	M	d_H
[5,2,5] Repetition code	5ℓ	2	3ℓ
[7,4,3] Hamming code	7ℓ	16	2ℓ
[8,4,2] Reed-Muller code	8ℓ	256	ℓ
(15, 256, 5) Nordstrom-Robinson code	15ℓ	256	3ℓ
[23,12,7] Golay code	23ℓ	4096	4ℓ
[24,12,8] Extended golay code	24ℓ	4096	4ℓ

Table 5 List of all encoded DNA strings for $(\mathbf{x}, \mathbf{y}) = (ATA, CGC)$ from [7, 4, 3] binary Hamming code. The code rate of the DNA code is 0.1904

[7, 4, 3] Hamming code	Encoded DNA code
0000000	$ATACGCATACGCATACGCATA$
1110000	$TATCGCTATGCGTATGCGTAT$
1001100	$TATGCGTATCGCTATGCGTAT$
0111100	$ATAGCGATAGCGATACGCATA$
0101010	$ATAGCGTATCGCATAGCGTAT$
1011010	$TATGCGATAGCGTATCGCATA$
1100110	$TATCGCATACGCTATCGCATA$
0010110	$ATACGCTATGCGATAGCGTAT$
1101001	$TATCGCATAGCGTATGCGATA$
0011001	$ATACGCTATCGCATACGCTAT$
0100101	$ATAGCGTATGCGATACGCTAT$
1010101	$TATGCGATACGCTATGCGATA$
1000011	$TATGCGTATGCGTATCGCTAT$
0110011	$ATAGCGATACGCATAGCGATA$
0001111	$ATACGCATAGCGATAGCGATA$
1111111	$TATCGCTATCGCTATCGCTAT$

For $\ell = 2$, one can obtain Non-Homopolymer map f of order 2 is given in Table 3(b) by considering $\mathbf{x} = CG$ and $\mathbf{y} = AT$. One can read the table as $f(1, CG) = TA$, $f(1, TA) = GC$ and $f(0, GC) = TA$ etc. Using the Non-Homopolymer map of order 2 (Table 3(b)), a binary string can also be encoded into a DNA string. Formally, the encoding using the Non-Homopolymer map is as follows.

Encoding 1 For positive integers n and ℓ, consider a mapping f as defined in the Definition 4. A binary string $\mathbf{a} = (a_1\ a_2 \ldots a_n) \in \{0, 1\}^n$ is encoded into $\mathbf{u} = (\mathbf{u}_1\ \mathbf{u}_2 \ldots \mathbf{u}_n) \in \{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\}^n$, in such a way that $\mathbf{u}_i = f(a_i, \mathbf{u}_{i-1})$ for each $i = 2, 3, \ldots, n$ and $\mathbf{u}_1 = h(a_1)$ where, $h : \{0, 1\} \to \{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\}$ such that $h(0) = h(1)^c$. Note that $\mathbf{u}_1 \in \{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\}$ initiates the encoding of the binary string, and therefore the length of the encoded DNA string $\mathbf{u}$ is $n\ell$. Clearly, the encoding of two distinct binary strings are always distinct. Observe that for $i = 2, 3, \ldots, n$, $f(a_i, \mathbf{u}_{i-1})^c = f(a_i, \mathbf{u}_{i-1}^c) = f(\bar{a}_i, \mathbf{u}_{i-1})$ where, $\bar{a}_i$ is binary complement of a_i.

Example 1 Consider a binary string $\mathbf{a} = (0\ 1\ 1\ 0)$. Let the encoding be initialized with $\mathbf{x} = CG$. So, one can encode the binary string $\mathbf{a}$ to a DNA string $(\mathbf{x}\ f(1, \mathbf{x})\ f(1, f(1, \mathbf{x}))\ f(0, f(1, f(1, \mathbf{x})))) = \mathbf{x}\mathbf{y}^c\mathbf{x}\mathbf{y} = CGTAGCTA$.

For positive integers n and ℓ $(< n)$, consider a subset $S \subseteq \{0, 1\}^n$. Each binary string from the set S is encoded using the Non-Homopolymer map, and the set of all encoded DNA strings is denoted by $f(S)$. In particular, let $\mathcal{B} = f(\{0, 1\}^n) \subseteq \Sigma_{DNA}^{n\ell}$ be the set of all possible DNA strings of length $n\ell$ and they are obtained by the encoding using Non-Homopolymer map.

Table 6 Codewords (each of length n) for $\lfloor n/2 \rfloor$ conflict free DNA codes with Hamming, reverse, reverse-complement and GC-content constraints. All the codes have code size improved from existing literature

Codewords for 2 conflict free DNA code with parameters (4, 12, 3)
$ACTG$, $AGCT$, $ATGC$, $CAGT$, $CGTA$, $CTAG$,
$GATC$, $GCAT$, $GTCA$, $TACG$, $TCGA$, $TGAC$
Codewords for 3 conflict free DNA code with parameters (6, 20, 4)
$ACAGTG$, $ACGTGA$, $AGCTAG$, $AGTGCA$, $ATCAGC$,
$CACTGT$, $CATGTC$, $CGACTA$, $CTAGCT$, $CTGTAC$,
$GACATG$, $GATCGA$, $GCTGAT$, $GTACAG$, $GTGACA$,
$TAGTCG$, $TCACGT$, $TCGATC$, $TGCACT$, $TGTCAC$
Codewords for 4 conflict free DNA code with parameters (8, 12, 6)
$ACAGATCG$, $AGCTACTC$, $CATACGTC$, $CGATCTGT$,
$CTCATCGA$, $CTGCATAC$, $GACGTATG$, $GAGTAGCT$,
$GCTAGACA$, $GTATGCAG$, $TCGATGAG$, $TGTCTAGC$
Codewords for 4 conflict free DNA code with parameters (9, 16, 6)
$ACAGTAGCT$, $AGCTACTGT$, $AGTAGCATC$, $ATACAGACG$,
$ATGATCGAG$, $CGTCTGTAT$, $CTACGATGA$, $CTCGATCAT$,
$GAGCTAGTA$, $GATGCTACT$, $GCAGACATA$, $TACTAGCTC$,
$TATGTCTGC$, $TCATCGTAG$, $TCGATGACA$, $TGTCATCGA$
Codewords for 4 conflict free DNA code with parameters (9, 2, 9)
$ACGATAGCA$, $TGCTATCGT$
Codewords for 5 conflict free DNA code with parameters (10, 16, 7)
$ACGTAGCAGA$, $ACTACAGACG$, $AGACGATGCA$,
$AGCGACTATC$, $ATAGCTCGTG$, $CACGAGCTAT$,
$CGTCTGTAGT$, $CTATCAGCGA$, $GATAGTCGCT$,
$GCAGACATCA$, $GTGCTCGATA$, $TATCGAGCAC$,
$TCGCTGATAG$, $TCTGCTACGT$, $TGATGTCTGC$,
$TGCATCGTCT$
Codewords for 5 conflict free DNA code with parameters (10, 8, 8)
$ACATGCGATC$, $CAGATACAGC$, $CGACATAGAC$,
$GATCGCATGT$, $GCTGTATCTG$, $GTCTATGTCG$,
$CTAGCGTACA$, $TGTACGCTAG$

$$\mathcal{B} = \begin{cases} (\{\mathbf{u}_1, \mathbf{u}_1^c\} \times \{\mathbf{u}_2, \mathbf{u}_2^c\})^{\frac{n-1}{2}} \times \{\mathbf{u}_1, \mathbf{u}_1^c\} & n \text{ is positive odd integer, and} \\ (\{\mathbf{u}_1, \mathbf{u}_1^c\} \times \{\mathbf{u}_2, \mathbf{u}_2^c\})^{\frac{n}{2}} & n \text{ is positive even integer} \end{cases}$$

where, $\{\mathbf{u}_1, \mathbf{u}_1^c\}$ and $\{\mathbf{u}_2, \mathbf{u}_2^c\}$ are pairwise disjoint and $\{\mathbf{u}_1, \mathbf{u}_1^c\} \cup \{\mathbf{u}_2, \mathbf{u}_2^c\} = \{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\}$. Note that, in the Example 1, the encoded DNA string $CGTAGCAT$ is 4 conflict free and also free from secondary structures. The GC-content of the DNA string is 4 ($= \lfloor n/2 \rfloor$). Therefore, in the following results, we obtained constraints on seed blocks ($\mathbf{x}$ and $\mathbf{y}$) of the Non-Homopolymer mapping so that the encoded DNA strings from binary strings satisfy those additional constraints.

Theorem 2 *For positive integers n ($n \geq 2$), ℓ, t and m ($\leq t\ell$), if $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$ such that each DNA string in the set $S = \{(\boldsymbol{x}\,\boldsymbol{y}_1\,\boldsymbol{x}_2\,\boldsymbol{y}_2 \ldots \boldsymbol{x}_t\,\boldsymbol{y}_t), (\boldsymbol{y}\,\boldsymbol{x}_1\,\boldsymbol{y}_2\,\boldsymbol{x}_2 \ldots$*

$\boldsymbol{y}_t\ \boldsymbol{x}_t)$: $\boldsymbol{x}_i \in \{\boldsymbol{x}, \boldsymbol{x}^c\}$ and $\boldsymbol{y}_i \in \{\boldsymbol{y}, \boldsymbol{y}^c\}$ for $i = 1, 2, \ldots, t\}$ is m conflict free then any binary string from $\{0, 1\}^n$ will be encoded into an m conflict free DNA string using Non-Homopolymer map.

Proof From Remark 7, if any DNA string in the set S is m conflict free then all the DNA sub-strings $(\mathbf{x}_1\ \mathbf{y}_1\ \mathbf{x}_2\ \mathbf{y}_2 \ldots \mathbf{x}_t\ \mathbf{y}_t)$ and $(\mathbf{y}_1\ \mathbf{x}_1\ \mathbf{y}_2\ \mathbf{x}_2 \ldots \mathbf{y}_t\ \mathbf{x}_t)$ are also m conflict free. Therefore, in the encoded DNA string $\mathbf{u} = (\mathbf{u}_1\ \mathbf{u}_2 \ldots$
$\mathbf{u}_n)$ (using Non-Homopolymer map), for $1 \leq i \leq n - 2t + 1$, consider $(\mathbf{u}_i\ \mathbf{u}_{i+1}$
$\mathbf{u}_{i+2} \ldots \mathbf{u}_{i+2t})$, which is also m conflict free where, $\mathbf{u}_j \in \{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\}$ for $j = i, i + 1, \ldots, i + 2t$. From Definition 2, $m < t\ell$, the result follows. □

For the various values of m and t in Theorem 2, one can observe the following two propositions:

Proposition 1 *For a positive integer ℓ, if $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$ such that each of the DNA strings $(\boldsymbol{x}\ \boldsymbol{y})$, $(\boldsymbol{x}\ \boldsymbol{y}^c)$, $(\boldsymbol{y}\ \boldsymbol{x})$ and $(\boldsymbol{y}\ \boldsymbol{x}^c)$ is ℓ conflict free then any binary string will be encoded into a ℓ conflict free DNA string using Non-Homopolymer map.*

Theorem 3 *For a positive integer ℓ (≥ 2), if $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$ such that each of the DNA string from the set $\{(\boldsymbol{x}\ \boldsymbol{y}_1\ \boldsymbol{x}_1\ \boldsymbol{y}_2\ \boldsymbol{x}_2) : \boldsymbol{x}_1, \boldsymbol{x}_2 \in \{\boldsymbol{x}, \boldsymbol{x}^c\}$ and $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \{\boldsymbol{y}, \boldsymbol{y}^c\}\}$ is free from reverse-complement sub-string(s) of length 3 then the encoded DNA string using Non-Homopolymer map is free from secondary structure of stem length 3.*

Proof For ℓ (≥ 2) and $\mathbf{x}, \mathbf{y} \in \Sigma_{DNA}^{\ell}$, consider a DNA string $(\mathbf{x}\ \mathbf{y}\ \mathbf{x}\ \mathbf{y}\ \mathbf{x})$ free from reverse-complement sub-strings of length 3. From Remark 5, the DNA strings $(\mathbf{x}^c\ \mathbf{y}^c\ \mathbf{x}^c\ \mathbf{y}^c\ \mathbf{x}^c)$ will also be free from reverse-complement sub-strings of length 3. Similarly remaining all DNA strings of 5 seed blocks are also free from reverse-complement sub-strings. Hence, from Definition 1 and Non-Homopolymer map, the encoded DNA string obtained from any binary string will be free from secondary structure of stem length 3. □

From Remark 4, if a DNA string of length n is free from reverse-complement sub-string of length m then the DNA string is also free from any reverse-complement sub-string of length t $(m \leq t \leq n)$. Therefore, one can observe following two propositions from Theorem 3.

Proposition 2 *For a positive integer ℓ (≥ 2), if $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$ such that each DNA string in the set $\{(\boldsymbol{x}\ \boldsymbol{y}^*\ \boldsymbol{x}^*\ \boldsymbol{y}^*), (\boldsymbol{y}\ \boldsymbol{x}^*\ \boldsymbol{y}^*\ \boldsymbol{x}^*) : \boldsymbol{x}^* \in \{\boldsymbol{x}, \boldsymbol{x}^c\}$ and $\boldsymbol{y}^* \in \{\boldsymbol{y},\ \ \boldsymbol{y}^c\}\}$ is free from secondary structures of stem length 3 then the encoded DNA string using Non-Homopolymer map is also free from secondary structure of stem length 3.*

Recall that any DNA string of length n is $\lfloor n/2 \rfloor$ conflict free then the DNA string is free from any consecutive repetitions of sub-strings of any length. Following theorem gives the condition on binary strings such that the encoded DNA string of length $n\ell$ is $\lfloor n\ell/2 \rfloor$ conflict free.

Theorem 4 *For positive integers n and ℓ, consider $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$ such that the DNA strings $(\boldsymbol{x}\ \boldsymbol{y})$, $(\boldsymbol{x}\ \boldsymbol{y}^c)$, $(\boldsymbol{y}\ \boldsymbol{x})$ and $(\boldsymbol{y}\ \boldsymbol{x}^c)$ are ℓ conflict free. If $\boldsymbol{a} = (a_1\ a_2 \ldots a_n)$ is a binary string such that $2\mu < \sum_{i=\lambda+1}^{\lambda+2\mu}(a_i a_{2\mu+i} + \bar{a}_i \bar{a}_{2\mu+i})$ for each positive even integer 2μ from the set*

*$\{1, 2, \ldots, \lfloor n/2 \rfloor\}$ and $\lambda = 0, 1, \ldots, n - 2\mu$ then the binary string **a** will be encoded (using Non-Homopolymer map) into a $\lfloor n\ell/2 \rfloor$ conflict free DNA string of length $n\ell$.*

Proof Consider a binary string $\mathbf{a} = (a_1\ a_2 \ldots a_n)$ which is encoded into DNA string $\mathbf{u} = (u_1\ u_2 \ldots u_n)$ using Non-Homopolymer map. For each positive even integer 2μ from the set $\{1, 2, \ldots, \lfloor n/2 \rfloor\}$, the DNA block $u_{2\mu+i} \in \{u_i, u_i^c\}$. For any binary symbol $a_i, a_{2\mu+i} \in \{0, 1\}$,

$$(a_i a_{2\mu+i} + \bar{a}_i \bar{a}_{2\mu+i}) = \begin{cases} 1 & \text{if } a_i = a_{2\mu+i} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\sum_{i=1}^{2\mu}(a_i a_{2\mu+i} + \bar{a}_i \bar{a}_{2\mu+i}) = 2\mu$ if and only if $a_i = a_{2\mu+i}$ for each $i = 1, 2, \ldots, 2\mu$. If the origin is shifted with λ, and $2\mu < \sum_{i=\lambda+1}^{\lambda+2\mu}(a_i a_{2\mu+i} + \bar{a}_i \bar{a}_{2\mu+i})$ for each λ and μ, then there is not exist the consecutive identical sub-string of length 2μ in the binary string. From Non-Homopolymer map, for a binary string, two consecutive sub-strings of odd length cannot be encoded into identical DNA sub-strings. Therefore, from Non-Homopolymer map and Proposition 1, the encoded DNA string will be a $\lfloor n\ell/2 \rfloor$ conflict free. □

The GC-content of encoded DNA string is calculated in the following lemma.

Lemma 3 *For positive integers n and ℓ, consider $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$ with GC-content $g_{\boldsymbol{x}}$ and $g_{\boldsymbol{y}}$. For the encoded DNA string $\boldsymbol{u} = (\boldsymbol{u}_1\ \boldsymbol{u}_2 \ldots \boldsymbol{u}_n) \in \{\boldsymbol{x}, \boldsymbol{x}^c, \boldsymbol{y},\ \boldsymbol{y}^c\}^n$ using Non-Homopolymer map, the GC-content of $\boldsymbol{u}$ will be*

$$g_{\boldsymbol{u}} = \begin{cases} g_{\boldsymbol{u}_1} + (g_{\boldsymbol{u}_1} + g_{\boldsymbol{u}_2})(n-1)/2 & \textit{if } n \textit{ is odd integer,} \\ (g_{\boldsymbol{u}_1} + g_{\boldsymbol{u}_2})n/2 & \textit{if } n \textit{ is even integer.} \end{cases}$$

Proof For positive integers n and ℓ $(< n)$, let a binary string $\mathbf{a} = (a_1\ a_2 \ldots a_n) \in \{0, 1\}^n$ be encoded into some $\mathbf{u} = (\mathbf{u}_1\ \mathbf{u}_2 \ldots \mathbf{u}_n) \in \{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\}^n$ using Non-Homopolymer map. In the Non-Homopolymer map, if $\mathbf{u}_1 \in \{\mathbf{x}, \mathbf{x}^c\}$ then DNA blocks $\mathbf{u}_{2j} \in \{\mathbf{y}, \mathbf{y}^c\}$ and $\mathbf{u}_{2j+1} \in \{\mathbf{x}, \mathbf{x}^c\}$, for $1 \leq 2j, 2j+1 \leq n$. Since, the GC-content of a DNA string and its complement DNA string are the same, the GC-content of each sub-string $(\mathbf{u}_{2j}\ \mathbf{u}_{2j+1})$ is $g_{\mathbf{x}} + g_{\mathbf{y}}$. Hence, if n is even, the GC-content of the encoded DNA string $\mathbf{u}$ is $g_{\mathbf{u}} = (g_{\mathbf{x}} + g_{\mathbf{y}})n/2$ and if n is odd then $g_{\mathbf{u}} = g_{\mathbf{x}} + (g_{\mathbf{x}} + g_{\mathbf{y}})(n-1)/2$. Similarly, if $\mathbf{u}_1 \in \{\mathbf{y}, \mathbf{y}^c\}$ then $g_{\mathbf{u}} = (g_{\mathbf{x}} + g_{\mathbf{y}})n/2$ for even n and $g_{\mathbf{u}} = g_{\mathbf{y}} + (g_{\mathbf{x}} + g_{\mathbf{y}})(n-1)/2$ when n is odd. Hence, the lemma follows. □

From Lemma 3, one can observe the following two propositions.

Proposition 3 *In Lemma 3, if $g_{\boldsymbol{x}} + g_{\boldsymbol{y}} = \ell$ then*

$$g_{\boldsymbol{u}} = \begin{cases} g_{\boldsymbol{u}_1} + \ell(n-1)/2 & \textit{if } n \textit{ is odd integer} \\ \ell n/2 & \textit{if } n \textit{ is even integer.} \end{cases}$$

The following proposition ensures that the GC-content of the encoded DNA string (using Non-Homopolymer map) is almost 50% of the length.

Proposition 4 *For positive integer n and ℓ, let $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$. If the GC-content of $\boldsymbol{x}$ and $\boldsymbol{y}$ are $\lfloor \ell/2 \rfloor$ and $\lceil \ell/2 \rceil$ respectively, then the GC-content of any encoded DNA string $\boldsymbol{u}$ (using*

Non-Homopolymer map) of length $n\ell$ is

$$g_{\boldsymbol{u}} = \begin{cases} \lfloor n\ell/2 \rfloor & \text{if } \boldsymbol{u}_1 \in \{\boldsymbol{x}, \boldsymbol{x}^c\}, \text{ and} \\ \lceil n\ell/2 \rceil & \text{if } \boldsymbol{u}_1 \in \{\boldsymbol{y}, \boldsymbol{y}^c\}. \end{cases}$$

Theorem 5 *For positive integers n and ℓ, let $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$. Using Non-Homopolymer map, if a binary string $(a_1\ a_2 \ldots a_n)$ is encoded into some $\boldsymbol{u} \in \{\boldsymbol{x}, \boldsymbol{x}^c, \boldsymbol{y}, \boldsymbol{y}^c\}^n$ then the binary string $(\bar{a}_1\ a_2 \ldots a_n)$ is encoded into $\boldsymbol{u}^c$ where, $\bar{a}_1$ is the binary complement of a_1.*

Proof The proof is done using induction on the index i $(i = 1, 2, \ldots, n)$. Now from the Non-Homopolymer map, $f(0, \mathbf{z})^c = f(0, \mathbf{z}^c)$ and $f(1, \mathbf{z})^c = f(1, \mathbf{z}^c)$, for each $\mathbf{z} \in \{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\}$. Consider the binary strings $(a_1\ a_2\ a_3 \ldots a_n)$ and $(\bar{a}_1\ a_2\ a_3 \ldots a_n)$, that are encoded into some DNA strings $(\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3 \ldots \mathbf{u}_n)$ and $(\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3 \ldots \mathbf{v}_n)$. From the Non-Homopolymer map, $\mathbf{u}_1^c = h(a_1)^c = h(\bar{a}_1) = \mathbf{v}_1$. Let $\mathbf{u}_i^c = \mathbf{v}_i$, for some $i \in \{1, 2, \ldots, n\}$. Consider $\mathbf{u}_{i+1}^c = f(a_{i+1}, \mathbf{u}_i)^c = f(a_{i+1}, \mathbf{u}_i^c) = f(a_{i+1}, \mathbf{v}_i) = \mathbf{v}_{i+1}$. Therefore, from induction, the binary strings $(a_1\ a_2\ a_3 \ldots a_n)$ and $(\bar{a}_1\ a_2\ a_3 \ldots a_n)$ are encoded into DNA strings which are complement to each other. □

In the following two theorems, the Hamming distance between the two DNA strings is calculated for binary strings with Hamming distance 1 and 2.

Theorem 6 *For positive integers n and ℓ, consider the binary strings $\boldsymbol{a} = (a_1\ a_2 \ldots a_n)$ and $\boldsymbol{b} = (a_1\ a_2 \ldots a_{i-1}\ \bar{a}_i\quad a_{i+1} \ldots a_n)$ $(1 \le i \le n)$, that are encoded into DNA strings $\boldsymbol{u} = (\boldsymbol{u}_1\ \boldsymbol{u}_2 \ldots \boldsymbol{u}_n)$ and $\boldsymbol{v} = (\boldsymbol{v}_1\ \boldsymbol{v}_2 \ldots \boldsymbol{v}_n)$ where, $\bar{a}_i$ is binary complement of a_i. Then, $d_H(\boldsymbol{u}, \boldsymbol{v}) = \ell(n - i + 1)$.*

Proof Consider the binary strings $\mathbf{a} = (a_1\ a_2 \ldots a_{i-1}\ a_i\ a_{i+1} \ldots a_n)$ and $\mathbf{b} = (a_1\ a_2 \ldots a_{i-1}\ \bar{a}_i\ a_{i+1} \ldots a_n)$ which can be encoded into $\mathbf{u} = (\mathbf{u}_1\ \mathbf{u}_2 \ldots \mathbf{u}_{i-1}\ \mathbf{u}_i$
$\mathbf{u}_{i+1} \ldots \mathbf{u}_n)$ and $\mathbf{v} = (\mathbf{v}_1\ \mathbf{v}_2 \ldots \mathbf{v}_{i-1}\ \mathbf{v}_i\ \mathbf{v}_{i+1} \ldots \mathbf{v}_n)$ respectively. From Non-Homopolymer map, $\mathbf{v}_j = \mathbf{u}_j$ $(j = 1, 2, \ldots i - 1)$ and $\mathbf{v}_j = \mathbf{u}_j^c$ $(j = i, i + 1 \ldots, n)$. Therefore, $d_H(\mathbf{u}, \mathbf{v}) = \ell(n - i + 1)$, since $d_H(\mathbf{x}, \mathbf{x}^c) = d_H(\mathbf{y}, \mathbf{y}^c) = \ell$. □

Theorem 7 *For positive integers n and ℓ, consider the binary strings $\boldsymbol{a} = (a_1\ a_2 \ldots a_n)$ and $\boldsymbol{b} = (a_1\ a_2 \ldots a_{i-1}\ \bar{a}_i\quad a_{i+1} \ldots a_{j-1}\ \bar{a}_j\ a_{j+1} \ldots a_n\,)$ $(1 \le i < j \le n)$ that are encoded into DNA strings $\boldsymbol{u} = (\boldsymbol{u}_1\ \boldsymbol{u}_2 \ldots \boldsymbol{u}_n)$ and $\boldsymbol{v} = (\boldsymbol{v}_1\ \boldsymbol{v}_2 \ldots \boldsymbol{v}_n)$ where, $\bar{a}_i$ is binary complement of a_i. Then, $d_H(\boldsymbol{u}, \boldsymbol{v}) = \ell(j - i)$.*

Proof The proof is similar to the Theorem 6 and follows from the fact that for any DNA string $\mathbf{x}$, $(\mathbf{x}^c)^c = \mathbf{x}$. □

The following theorem provides a bound on the Hamming distance on the encoded DNA strings.

Theorem 8 *For positive integers n, ℓ and σ $(\le \ell)$, consider $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$ and $\sigma = \min\{d_H(z_1, z_2),\ \ n - d_H(z_1, z_2) : z_1 \in \{\boldsymbol{x}, \boldsymbol{x}^c\}$ and $z_2 \in \{\boldsymbol{y}, \boldsymbol{y}^c\}\}$. Let binary strings $\boldsymbol{a}$, $\boldsymbol{b} \in \{0, 1\}^n$ be encoded into the DNA strings $\boldsymbol{u} = (\boldsymbol{u}_1\ \boldsymbol{u}_2 \ldots \boldsymbol{u}_n)$ and $\boldsymbol{v} = (\boldsymbol{v}_1\ \boldsymbol{v}_2 \ldots \boldsymbol{v}_n)$. For some $a, b \in \{0, 1\}$, if $\boldsymbol{u}' = (\boldsymbol{u}\ f(a, \boldsymbol{u}_n))$ and $\boldsymbol{v}' = (\boldsymbol{v}\ f(b, \boldsymbol{v}_n))$ then*

$$d_H(\boldsymbol{u}', \boldsymbol{v}') \ge \begin{cases} \sigma(d_H(\boldsymbol{a}, \boldsymbol{b}) + d_H(a, b)) & \text{if } d_H(\boldsymbol{a}, \boldsymbol{b}) \text{ is even} \\ \sigma(d_H(\boldsymbol{a}, \boldsymbol{b}) - d_H(a, b) + 1) & \text{if } d_H(\boldsymbol{a}, \boldsymbol{b}) \text{ is odd.} \end{cases}$$

Proof The proof follows from the following two facts. (1) For a positive integer m and any $\mathbf{x}, \mathbf{y} \in \Sigma_{DNA}^m$, if $d_H(\mathbf{x}, \mathbf{y}) = t$ then $d_H(\mathbf{x}^c, \mathbf{y}) \geq m - t$, and (2) For $\mathbf{z} \in \{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\}$ and $a, b \in \{0, 1\}$, $f(c, \mathbf{z})^c = f(c, \mathbf{z}^c) = f(\bar{c}, \mathbf{z})$. So we can derive,

$$d_H(\mathbf{u}', \mathbf{v}') \geq \begin{cases} \sigma(d_H(\mathbf{a}, \mathbf{b}) + 1) & \text{if } d_H(\mathbf{a}, \mathbf{b}) \text{ is even and } a \neq b \\ \sigma d_H(\mathbf{a}, \mathbf{b}) & \text{if } d_H(\mathbf{a}, \mathbf{b}) \text{ is even and } a = b \\ \sigma d_H(\mathbf{a}, \mathbf{b}) & \text{if } d_H(\mathbf{a}, \mathbf{b}) \text{ is odd and } a \neq b \\ \sigma(d_H(\mathbf{a}, \mathbf{b}) + 1) & \text{if } d_H(\mathbf{a}, \mathbf{b}) \text{ is odd and } a = b. \end{cases}$$

Hence the result follows. □

Motivated by Theorem 8, one can observe the following proposition.

Proposition 5 *For positive integers n and ℓ, consider $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$ and $\ell = d_H(z_1, z_2)$, for each $z_1 \in \{\boldsymbol{x}, \boldsymbol{x}^c\}$ and $z_2 \in \{\boldsymbol{y}, \boldsymbol{y}^c\}$. Let two binary strings $\boldsymbol{a}$, $\boldsymbol{b} \in \{0, 1\}^n$ be encoded into the DNA strings $\boldsymbol{u} = (\boldsymbol{u}_1\ \boldsymbol{u}_2 \ldots \boldsymbol{u}_n)$ and $\boldsymbol{v} = (\boldsymbol{v}_1\ \boldsymbol{v}_2 \ldots \boldsymbol{v}_n)$. For some $a, b \in \{0, 1\}$, if $\boldsymbol{u}' = (\boldsymbol{u}\ f(a, \boldsymbol{u}_n))$ and $\boldsymbol{v}' = (\boldsymbol{v}\ f(b, \boldsymbol{v}_n))$ then*

$$d_H(\boldsymbol{u}', \boldsymbol{v}') = \begin{cases} d_H(\boldsymbol{u}, \boldsymbol{v}) + \ell d_H(a, b) & \text{if } d_H(\boldsymbol{a}, \boldsymbol{b}) \text{ is even} \\ d_H(\boldsymbol{u}, \boldsymbol{v}) + \ell(1 - d_H(a, b)) & \text{if } d_H(\boldsymbol{a}, \boldsymbol{b}) \text{ is odd.} \end{cases}$$

In order to establish the proposed mapping as an isometry from the set of binary strings to the set of DNA strings where, Hamming distance is taken for the set of DNA strings, we introduce a new distance between two binary strings in the following definition.

Definition 5 Let $n(> 1)$ be an integer and Σ be an alphabet of size q (≤ 2). For any $\mathbf{a} = (a_1\ a_2 \ldots a_n)$, $\mathbf{b} = (b_1\ b_2 \ldots b_n) \in \Sigma^n$, let $P = \{i : a_i \neq b_i \text{ and } i \in \{1, 2, \ldots, n\}\}$ and

$$S = \begin{cases} P & \text{if } d_H(\mathbf{a}, \mathbf{b}) \text{ is even} \\ P \cup \{n + 1\} & \text{if } d_H(\mathbf{a}, \mathbf{b}) \text{ is odd.} \end{cases}$$

If $S \neq \emptyset$, we can denote $S = \{s_1, s_2, \ldots, s_{|S|}\}$ such that, for each $s_j < s_{j+1}$, $j = 1, 2, \ldots, |S| - 1$. We define a map $d_{NHo} : \Sigma^n \times \Sigma^n \to \mathbb{R}$ such thatl

$$d_{NHo}(\mathbf{a}, \mathbf{b}) = \begin{cases} \ell \sum_{i=1}^{|S|/2} (s_{2i} - s_{2i-1}) & \text{if } |S| > 0, \\ 0 & \text{if } |S| = 0, \end{cases}$$

where ℓ is a positive integer.

Remark 9 The map $d_{NHo} : \Sigma^n \times \Sigma^n \to \mathbb{R}$ is indeed a distance (called Non-Homopolymer distance), because it holds all the properties including triangle inequality of distance. The first three properties of metric are easy to prove and the triangle inequality property can be proved using induction on length n. For a code $\mathcal{C} \subseteq \Sigma^n$, the minimum distance $d_{NHo} = \min\{d_{NHo}(\mathbf{a}, \mathbf{b}) : \mathbf{a} \neq \mathbf{b}, \text{ and } \mathbf{a}, \mathbf{b} \in \mathcal{C}\}$.

For example, consider $n = 5$, $\ell = 2$ and $\Sigma = \{0, 1\}$. For $\mathbf{a} = (1\ 1\ 1\ 1\ 0)$ and $\mathbf{b} = (0\ 1\ 1\ 0\ 0)$, $d_{NHo}(\mathbf{a}, \mathbf{b}) = 6$ where, $S = \{1, 4\}$.

Theorem 9 *For positive integers n and ℓ, the Non-Homopolymer map is a distance preserving encoding between $(\{0,1\}^n, d_{NHo})$ and $(\mathcal{B}, d_H)$.*

Proof The theorem is proved using induction on the string length n. The base case, $n = 1$, is obvious from the Non-Homopolymer map and the Non-Homopolymer distance. For the inductive step, assume that the distance is preserved for $n = k$ where, $\mathbf{a} = (a_1\ a_2 \dots a_k)$ and $\mathbf{b} = (b_1\ b_2 \dots b_k)$ with the support set S, are encoded into DNA strings $\mathbf{u} = (\mathbf{u}_1\ \mathbf{u}_2 \dots \mathbf{u}_k)$ and $\mathbf{v} = (\mathbf{v}_1\ \mathbf{v}_2 \dots \mathbf{v}_k)$. To prove that the distance is preserved for $n = k+1$, consider $\mathbf{a}' = (\mathbf{a}\ a_{k+1}) = (a_1\ a_2 \dots a_k\ a_{k+1})$ and $\mathbf{b}' = (\mathbf{b}\ b_{k+1}) = (b_1\ b_2 \dots b_k\ b_{k+1})$ with the support set S' where, $a_{k+1}, b_{k+1} \in \{0,1\}$. Let the strings $\mathbf{a}$ and $\mathbf{b}$ be encoded into DNA strings $\mathbf{u}' = (\mathbf{u}_1\ \mathbf{u}_2 \dots \mathbf{u}_k\ \mathbf{u}_{k+1})$ and $\mathbf{v}' = (\mathbf{v}_1\ \mathbf{v}_2 \dots \mathbf{v}_k\ \mathbf{v}_{k+1})$ where, $\mathbf{u}_{k+1}, \mathbf{v}_{k+1} \in \{\mathbf{x}, \mathbf{y}, \mathbf{x}^c, \mathbf{y}^c\}$. Now, there are four cases. (i) Consider $d_H(\mathbf{a}, \mathbf{b})$ is even and $a_{k+1} = b_{k+1}$. Note that both $\mathbf{u}_k$ and $\mathbf{v}_k$ are member of either $\{\mathbf{x}, \mathbf{x}^c\}$ or $\{\mathbf{y}, \mathbf{y}^c\}$. On the other hand, $d_{NHo}(\mathbf{a}', \mathbf{b}') = d_{NHo}(\mathbf{a}, \mathbf{b})$, since $S' = S$. (ii) If $d_H(\mathbf{a}, \mathbf{b})$ is even and $a_{k+1} \neq b_{k+1}$ then $\mathbf{u}_{k+1} = \mathbf{v}^c_{k+1}$. So, for $\ell = d_H(\mathbf{x}, \mathbf{x}^c) = d_H(\mathbf{y}, \mathbf{y}^c)$, $d_H(\mathbf{u}', \mathbf{v}') = d_H(\mathbf{u}, \mathbf{v}) + \ell$ and $d_{NHo}(\mathbf{a}', \mathbf{b}') = d_{NHo}(\mathbf{a}, \mathbf{b}) + \ell$ since, $S' = \{k+1, k+2\} \cup S$. (iii) If $d_H(\mathbf{a}, \mathbf{b})$ is odd and $a_{k+1} = b_{k+1}$ then $d_H(\mathbf{u}', \mathbf{v}') = d_H(\mathbf{u}, \mathbf{v}) + m$ and $d_{NHo}(\mathbf{a}', \mathbf{b}') = d_{NHo}(\mathbf{a}, \mathbf{b}) + \ell$ since, $S' = \{k+2\} \cup S \backslash \{k+1\}$. (iv) If $d_H(\mathbf{a}, \mathbf{b})$ is odd and $a_{k+1} \neq b_{k+1}$ then $d_H(\mathbf{u}', \mathbf{v}') = d_H(\mathbf{u}, \mathbf{v})$ and $d_{NHo}(\mathbf{a}', \mathbf{b}') = d_{NHo}(\mathbf{a}, \mathbf{b})$ since, $S' = S$. Hence, the result follows from Proposition 5. □

Theorem 10 *For a binary code $\mathcal{C}(n, M, d_{NHo})$ where, d_{NHo} is the Non-Homopolymer distance (Definition 5), there exists a DNA code $f(\mathcal{C})$ with codeword length $n\ell$, size M and the minimum Hamming distance $d_H = d_{NHo}$.*

Proof The proof follows from the Non-Homopolymer map and Theorem 9. □

Theorem 11 *For any binary code $\mathcal{C}$ with the minimum distance $d_{NHo} \leq n\ell/2$ ($n, \ell \in \mathbb{Z}^+$), there exists a DNA code $f(\mathcal{C})(n\ell, M, d_H)$ with complement constraint.*

Proof Let binary strings $\mathbf{a}$ and $\mathbf{b}$ of length n be encoded into DNA strings $\mathbf{u}$ and $\mathbf{v}$ of length $n\ell$. From the property of the complement of a DNA string, $d_H(\mathbf{u}, \mathbf{v}^c) \geq n\ell - d_H(\mathbf{u}, \mathbf{v})$. From Theorem 9, $d_H(\mathbf{u}, \mathbf{v}) = d_{NHo}(\mathbf{a}, \mathbf{b}) \leq n\ell/2$. Therefore, $d_H(\mathbf{u}, \mathbf{v}^c) \geq n\ell/2$ and hence, the result follows. □

Theorem 12 *For a positive integer n, if a binary linear code with codeword length n contains $(1\ 0\ 0 \dots 0)$ as a codeword, then the encoded DNA code (using Non-Homopolymer map) will satisfy complement constraint.*

Proof Consider a binary linear code containing the codeword $(1\ 0\ 0 \dots 0)$ of length n. For any codeword $\mathbf{a} = (a_1\ a_2 \dots a_n)$ of the binary linear code, $(1\ 0\ 0 \dots 0) + (a_1\ a_2 \dots a_n) = (\bar{a}_1\ a_2 \dots a_n) = \mathbf{b}$ is also a codeword of the code. Therefore, from Theorem 5, for each binary codeword $\mathbf{a}$, there exists a binary codeword $\mathbf{b}$ such that the encoded DNA strings from $\mathbf{a}$ and $\mathbf{b}$ will be complement to each other. Hence, by the distance property, the theorem is proved. □

Theorem 13 *For positive integers n and ℓ, let $\boldsymbol{x}, \boldsymbol{y} \in \Sigma^{\ell}_{DNA}$. Then, for any encoded DNA strings $\boldsymbol{u}, \boldsymbol{v} \in f(\{0,1\}^n)$ using Non-Homopolymer map,*

$$d_H(\boldsymbol{u}, \boldsymbol{v}^r) \geq \begin{cases} n \min\{d_H(\boldsymbol{x}, \boldsymbol{y}^r), d_H(\boldsymbol{x}, \boldsymbol{y}^{rc})\}, & \text{if } n \text{ is even,} \\ \min\{d_H(\boldsymbol{x}, \boldsymbol{x}^r), d_H(\boldsymbol{y}, \boldsymbol{y}^r), d_H(\boldsymbol{x}, \boldsymbol{x}^{rc}), d_H(\boldsymbol{y}, \boldsymbol{y}^{rc})\}, & \text{if } n \text{ is odd.} \end{cases}$$

Proof For $\mathbf{x}, \mathbf{y} \in \Sigma_{DNA}^{\ell}$, let the binary strings $\mathbf{a}, \mathbf{b} \in \{0, 1\}^n$ of length n be encoded into DNA strings $\mathbf{u} = (\mathbf{u}_1\ \mathbf{u}_2 \ldots \mathbf{u}_n)$, $\mathbf{v} = (\mathbf{v}_1\ \mathbf{v}_2 \ldots \mathbf{v}_n)$ in $\{\mathbf{x}, \mathbf{x}^c, \mathbf{y}, \mathbf{y}^c\}$, where $\mathbf{u}_{2i}, \mathbf{v}_{2i} \in \{\mathbf{u}_2, \mathbf{u}_2^c\}$ and $\mathbf{u}_{2i+1}, \mathbf{v}_{2i+1} \in \{\mathbf{u}_1, \mathbf{u}_1^c\}$ for $1 \leq 2i, 2i+1 \leq n$. The set $f(\{0, 1\}^n)$ is the collection of all possible DNA strings such that obtained DNA blocks will be from $\{\mathbf{u}_2, \mathbf{u}_2^c\}$ and $\{\mathbf{u}_1, \mathbf{u}_1^c\}$ at even positions and odd positions respectively. Consider $d_H(\mathbf{u}, \mathbf{v}^r) = \sum_{j=1}^{n} d_H(\mathbf{u}_j, \mathbf{v}_{n-j+1}^r)$. Now two cases may arise.

Case 1: If n is odd, then j and $n - j + 1$ both are either even or odd. If both j and $n - j + 1$ are even then $\mathbf{u}_j, \mathbf{v}_{n-j+1} \in \{\mathbf{u}_2, \mathbf{u}_2^c\}$, and if both j and $n - j + 1$ are odd then $\mathbf{u}_j, \mathbf{v}_{n-j+1} \in \{\mathbf{u}_1, \mathbf{u}_1^c\}$. Therefore, $\mathbf{u}, \mathbf{v}^r \in f(\{0, 1\}^n)$ and, from the Non-Homopolymer map, $d_H(\mathbf{u}_j, \mathbf{v}_{n-j+1}^r) \geq \min\{d_H(\mathbf{x}, \mathbf{x}^r), d_H(\mathbf{y}, \mathbf{y}^r), d_H(\mathbf{x}, \mathbf{x}^{rc}), d_H(\mathbf{y}, \mathbf{y}^{rc})\}$.

Case 2: If n is even, then the parity j and $n - j + 1$ will be different. So, for even j, $\mathbf{u}_j \in \{\mathbf{u}_2, \mathbf{u}_2^c\}$ and $\mathbf{v}_{n-j+1} \in \{\mathbf{u}_1, \mathbf{u}_1^c\}$, and, for odd j, $\mathbf{u}_j \in \{\mathbf{u}_1, \mathbf{u}_1^c\}$ and $\mathbf{v}_{n-j+1} \in \{\mathbf{u}_2, \mathbf{u}_2^c\}$. Therefore, from Non-Homopolymer map and the fact that, for any $\mathbf{z}_1, \mathbf{z}_2 \in \Sigma_{DNA}^{\ell}$, $d_H(\mathbf{z}_1, \mathbf{z}_2^r) = d_H(\mathbf{z}_1^r, \mathbf{z}_2)$, we obtain $d_H(\mathbf{u}_j, \mathbf{v}_{n-j+1}^r) \geq \min\{d_H(\mathbf{x}, \mathbf{y}^r), d_H(\mathbf{x}, \mathbf{y}^{rc})\}$. Hence the result follows for every n. □

Similarly one can get result for reverse-complement constraint as given in the following preposition.

Proposition 6 *For positive integers n and ℓ, let $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$. Then, for any encoded DNA strings $\boldsymbol{u}, \boldsymbol{v} \in f(\{0, 1\}^n)$ using the Non-Homopolymer map,*

$$d_H(\boldsymbol{u}, \boldsymbol{v}^{rc}) \geq \begin{cases} n \min\{d_H(\boldsymbol{x}, \boldsymbol{y}^r), d_H(\boldsymbol{x}, \boldsymbol{y}^{rc})\}, & \text{if } n \text{ is even,} \\ \min\{d_H(\boldsymbol{x}, \boldsymbol{x}^r), d_H(\boldsymbol{y}, \boldsymbol{y}^r), d_H(\boldsymbol{x}, \boldsymbol{x}^{rc}), d_H(\boldsymbol{y}, \boldsymbol{y}^{rc})\}, & \text{if } n \text{ is odd.} \end{cases}$$

Theorem 14 *For an even positive integer n and a positive integer ℓ, consider $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_{DNA}^{\ell}$ such that $d_H(\boldsymbol{x}, \boldsymbol{y}^{rc}) = d_H(\boldsymbol{x}, \boldsymbol{y}^r) = \ell$. Then, the DNA codes constructed using the Non-Homopolymer map will satisfy the reverse constraint.*

Proof If $d_H(\mathbf{x}, \mathbf{y}^{rc}) = d_H(\mathbf{x}, \mathbf{y}^r) = \ell$ then, from Theorem 13, $d_H(\mathbf{u}, \mathbf{v}^r) \geq n \min\{d_H(\mathbf{x}, \mathbf{y}^r), d_H(\mathbf{x}, \mathbf{y}^{rc})\} = n\ell$. But the length of the encoded DNA string is $n\ell$ so, $d_H \leq n\ell$ and therefore, $d_H(\mathbf{u}, \mathbf{v}^r) \geq d_H$ for any DNA code constructed using the Non-Homopolymer map. □

Lemma 4 *For positive integers n and ℓ, if the binary strings $\boldsymbol{a}$ and $\boldsymbol{b}$ of length n are encoded into DNA strings $\boldsymbol{u}$ and $\boldsymbol{v}$ using the Non-Homopolymer map then*

$$n - \lfloor d_H(\boldsymbol{a}, \boldsymbol{b})/2 \rfloor \geq \frac{1}{\ell} d_H(\boldsymbol{u}, \boldsymbol{v}) \geq \lceil d_H(\boldsymbol{a}, \boldsymbol{b})/2 \rceil.$$

Proof For a positive integer n, if $S \subseteq \{1, 2, \ldots, n, n+1\}$ is a set with even cardinality such that, for each $s_j \in S$ $(j = 1, 2, \ldots, |S| - 1)$, $s_j < s_{j+1}$ then one can observe that:

$$n - \frac{|S|}{2} \geq \sum_{i=1}^{|S|/2} (s_{2i} - s_{2i-1}) \geq \frac{|S|}{2}.$$

From the Definition 5, $d_H(\mathbf{a}, \mathbf{b}) \in \{|S|, |S| - 1\}$, and therefore, the result follows. □

Theorem 15 *For positive integers n and ℓ, suppose a binary code exists with the minimum Hamming distance d_H and the minimum distance d_{NHo}. Then, $\ell(n - \lfloor d_H/2 \rfloor) \geq d_{NHo} \geq \ell\lceil d_H/2 \rceil$.*

Proof For any code $\mathcal{C}$ with the minimum Hamming distance d_H, if $\mathbf{a}, \mathbf{b} \in \mathcal{C}$ then $\lceil d_H(\mathbf{a}, \mathbf{b})/2 \rceil \geq \lceil d_H/2 \rceil$ and $n - \lfloor d_H/2 \rfloor \geq n - \lfloor d_H(\mathbf{a}, \mathbf{b})/2 \rfloor \geq \lceil d_H/2 \rceil$. The proof follows from Lemma 4 and Theorem 9. □

In the following theorem, a constraint on binary string is imposed in such a way that the encoded DNA string of length $n\ell$ will be $\lfloor n\ell/2 \rfloor$ conflict free.

Theorem 16 *For positive integers n and ℓ, and any positive even integer $2\mu \in \{1, 2, \ldots, \lfloor n/2 \rfloor\}$, consider a binary code with codeword length n, such that for each codeword $(a_1\ a_2 \ldots a_n)$,*

$$2\mu < \sum_{i=\lambda+1}^{\lambda+2\mu} (a_i a_{2\mu+i} + \bar{a}_i \bar{a}_{2\mu+i}),$$

where $\lambda = 1, 2, \ldots, n - 2\mu$. Then there exists a $\lfloor n\ell/2 \rfloor$ conflict free DNA code with codeword length $n\ell$.

Proof The proof follows from Definition 3 and Theorem 4. □

Lemma 5 *Consider two seed blocks $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ $i = 1, 2$ such that $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = d(\boldsymbol{x}_1, \boldsymbol{x}_2^c) = d(\boldsymbol{y}_1, \boldsymbol{y}_2) = d(\boldsymbol{y}_1, \boldsymbol{y}_2^c) = \ell$, and a binary code $\mathcal{C}$ with the parameter (n, M, d_{NHo}). The parameter of the DNA code $\mathcal{C}_{DNA} = \mathcal{C}_{DNA_1} \cup \mathcal{C}_{DNA_2}$ will be $(n\ell, 2M, d_H)$ where, the DNA code $\mathcal{C}_{DNA_i}$ is encoded from the binary code $\mathcal{C}$ using Non-Homopolymer map with the seed block $(\boldsymbol{x}_i, \boldsymbol{y}_i)$.*

Proof From Theorem 10, the parameters of the encoded DNA code $\mathcal{C}_{DNA_i}$ is $(n\ell, M, d_H)$ for $i = 1, 2$. The codeword length of both the encoded DNA codes $\mathcal{C}_{DNA_1}$ and $\mathcal{C}_{DNA_2}$ are same and equal to n. Therefore, the codeword length of the DNA code $\mathcal{C}_{DNA_1} \cup \mathcal{C}_{DNA_2}$ will also be n. The size of the DNA code $\mathcal{C}_{DNA_1} \cup \mathcal{C}_{DNA_2}$ will be not be more than $2M$. Let $\mathbf{a} = (a_1\ a_2 \ldots a_n)$ be a codeword in $\mathcal{C}$. Let the codeword $\mathbf{a}$ is encoded into $\mathbf{u} = (\mathbf{u}_1\ \mathbf{u}_2 \ldots \mathbf{u}_n) \in \mathcal{C}_{DNA_1}$ and $\mathbf{v} = (\mathbf{v}_1\ \mathbf{v}_2 \ldots \mathbf{v}_n) \in \mathcal{C}_{DNA_2}$ using seed blocks $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$. The encoded DNA strings $\mathbf{u}_1 \in \{\mathbf{x}_1, \mathbf{x}_1^c\}$, $\mathbf{v}_1 \in \{\mathbf{x}_2, \mathbf{x}_2^c\}$ where, $\mathbf{x}_1 \neq \mathbf{x}_2$ and $\mathbf{x}_2 \neq \mathbf{x}_2^c$. Therefore, $\mathbf{u} \neq \mathbf{v}$ for any case and hence, the encoded DNA strings are not identical. It follows that the size of the DNA code $\mathcal{C}_{DNA_1} \cup \mathcal{C}_{DNA_2}$ will be $2M$. In addition, if $d(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_1, \mathbf{x}_2^c) = d(\mathbf{y}_1, \mathbf{y}_2) = d(\mathbf{y}_1, \mathbf{y}_2^c) = \ell$ then $d_H(\mathbf{u}, \mathbf{v}) = n\ell$ $(\leq d_H)$ for any $\mathbf{u} \in \mathcal{C}_{DNA_1}$ and $\mathcal{C}_{DNA_2}$. Therefore, the minimum Hamming distance of the DNA code $\mathcal{C}_{DNA_1} \cup \mathcal{C}_{DNA_2}$ will be d_H. □

One can generalize Lemma 5 in the following proposition.

Proposition 7 *Consider r seed blocks $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ $i = 1, 2, \ldots, r$ such that $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = d(\boldsymbol{x}_i, \boldsymbol{x}_j^c) = d(\boldsymbol{y}_i, \boldsymbol{y}_j) = d(\boldsymbol{y}_i, \boldsymbol{y}_j^c) = \ell$ $(1 \leq i < j \leq r)$, and a binary code $\mathcal{C}$ with the parameter (n, M, d_{NHo}). The parameter of the DNA code $\mathcal{C}_{DNA} = \bigcup_{i=1}^{r} \mathcal{C}_{DNA_i}$ will be $(n\ell, rM, d_H)$ where, the DNA code $\mathcal{C}_{DNA_i}$ is encoded from the binary code $\mathcal{C}$ using the Non-Homopolymer map with the seed block $(\boldsymbol{x}_i, \boldsymbol{y}_i)$.*

One can observe the following remark from reverse, complement and reverse-complement DNA strings.

Remark 10 Using the Non-Homopolymer map, from some seed blocks $(\mathbf{x}, \mathbf{y})$, an ℓ conflict free DNA code is obtained with reverse, reverse-complement and GC-content constraints such that each DNA string is free from secondary structures if and only if an ℓ conflict free DNA code can also be obtained from seed blocks $(\mathbf{x}^r, \mathbf{y}^r)$ with reverse, reverse-complement and GC-content constraints such that each DNA string is free from secondary structures. The statement is also true for DNA codes generated from seed blocks $(\mathbf{x}^c, \mathbf{y}^c)$, $(\mathbf{x}^{rc}, \mathbf{y}^{rc})$, $(\mathbf{y}, \mathbf{x})$, $(\mathbf{y}^r, \mathbf{x}^r)$, $(\mathbf{y}^c, \mathbf{x}^c)$ and $(\mathbf{y}^{rc}, \mathbf{x}^{rc})$ using the Non-Homopolymer map.

Lemma 6 *Consider, for $i = 1, 2$, two seed blocks $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ such that $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = d(\boldsymbol{x}_1, \boldsymbol{x}_2^c) = d(\boldsymbol{y}_1, \boldsymbol{y}_2) = d(\boldsymbol{y}_1, \boldsymbol{y}_2^c) = \ell$, and two binary codes $\mathcal{C}_i$ with the parameter (n, M_i, d_{NHo_i}). The parameter of the DNA code $\mathcal{C}_{DNA} = \mathcal{C}_{DNA_1} \cup \mathcal{C}_{DNA_2}$ will be $(n\ell, M, d_H)$ where, the DNA code $\mathcal{C}_{DNA_i}$ is encoded from the binary code $\mathcal{C}_i$ using the Non-Homopolymer map with the seed block $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $M = M_1 + M_2$ and $d_H = \min\{d_{NHo_i} : i = 1, 2\}$.*

Proof The proof is similar to that of the proof of Lemma 5 and it follows from the definition of the minimum Hamming distance of a code. □

Proposition 8 *For $i = 1, 2, \ldots, r$, consider seed blocks $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ such that $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = d(\boldsymbol{x}_i, \boldsymbol{x}_j^c) = d(\boldsymbol{y}_i, \boldsymbol{y}_j) = d(\boldsymbol{y}_i, \boldsymbol{y}_j^c) = \ell$ $(1 \le i < j \le r)$, and binary codes $\mathcal{C}_i$ with the parameter (n, M_i, d_{NHo_i}). The parameter of the DNA code $\mathcal{C}_{DNA} = \bigcup_{i=1}^r \mathcal{C}_{DNA_i}$ will be $(n\ell, M, d_H)$ where, the DNA code $\mathcal{C}_{DNA_i}$ is encoded from the binary code $\mathcal{C}$ using Non-Homopolymer map with the seed block $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $M = \sum_{i=1}^r M_i$ and $d_H = \min\{d_{NHo_i} : i = 1, 2, \ldots, r\}$.*

Proposition 9 *As considered in Proposition 8, for a DNA code $\mathcal{C}_{DNA} = \bigcup_{i=1}^r \mathcal{C}_{DNA_i}$ with the parameter $(n\ell, \sum_{i=1}^r M_i, d_H)$, the code rate is:*

$$\mathcal{R} = \frac{log_4\left(\sum_{i=1}^r M_i\right)}{n\ell} \le \frac{r}{2\ell}.$$

The bound on code rate in the Proposition 9 is obtained by taking $M_i = 2^n$ for each $i = 1, 2, \ldots, r$.

Remark 11 For any ℓ conflict free DNA code $\mathcal{C}_{DNA}$ with the parameters $(n\ell, 4^n, d_H)$ and encoded from $\mathcal{C} = \mathbb{Z}_2^n$ using the Non-Homopolymer map, the code rate is $1/2\ell$. Note that for $\ell = 2$, the code rate is $1/4$. In addition, the DNA code satisfies all the constraints reverse, reverse-complement, GC-content. All the DNA codewords are also free from secondary like structures and homopolymers.

Lemma 7 *For $d_H < n\ell/2$, there exists DNA code $\mathcal{C}_{DNA}(n\ell, M, d_H)$ encoded from a binary code $\mathcal{C}(n, M, d_{NHo})$ such that*

$$M \ge \frac{2^n}{\sum_{i=0}^{\left\lceil \frac{2}{\ell} d_H \right\rceil - 1} \binom{n}{i}}. \tag{1}$$

Proof From Lemma 4, one can observe that $d_H(\mathbf{a}, \mathbf{b}) \leq \left\lceil \frac{2}{\ell} d_H(\mathbf{u}, \mathbf{v}) \right\rceil$ for any binary codewords $\mathbf{a}$ and $\mathbf{b}$ in $\mathcal{C}$ which can be encoded into DNA codewords $\mathbf{u}$ and $\mathbf{v}$ in $\mathcal{C}_{DNA}$. The bounds are true for any distinct $\mathbf{a}$ and $\mathbf{b}$ in $\mathcal{C}$ therefore, $d'_H \leq \left\lceil \frac{2}{\ell} d_H \right\rceil$ where, d'_H is the minimum Hamming distance for the binary code $\mathcal{C}$. From the Gilbert-Varshamov bound for binary code, one can obtain the bounds in (1) for $d_H < n\ell/2$. □

Now, one can easily obtain the following proposition.

Proposition 10 *If there exists an ℓ conflict free DNA code $\mathcal{C}_{DNA}$ with all the constraints reverse, reverse-complement, GC-content such that each DNA codeword is free from secondary like structures then*

$$A_4^{cf,hf,GC,r,rc}(n, d_H, \ell) \geq \frac{2^n}{\sum_{i=0}^{\left\lceil \frac{2}{\ell} d_H \right\rceil - 1} \binom{n}{i}}, \textit{ for } d_H < n\ell/2.$$

Lemma 8 *For $d_H \geq n\ell/2$, there exists DNA code $\mathcal{C}_{DNA}(n\ell, M, d_H)$ encoded from a binary code $\mathcal{C}(n, M, d_{NHo})$ such that*

$$M \geq \frac{2^n}{\sum_{i=0}^{2n - \left\lceil \frac{2}{\ell} d_H \right\rceil} \binom{n}{i}}. \tag{2}$$

Proof From Lemma 4, one can observe that $d_H(\mathbf{a}, \mathbf{b}) \leq 2n - \left\lceil \frac{2}{\ell} d_H(\mathbf{u}, \mathbf{v}) \right\rceil + 1$ for any binary codewords $\mathbf{a}$ and $\mathbf{b}$ in $\mathcal{C}$ which can be encoded into DNA codewords $\mathbf{u}$ and $\mathbf{v}$ in $\mathcal{C}_{DNA}$. The bounds are true for any distinct $\mathbf{a}$ and $\mathbf{b}$ in $\mathcal{C}$ therefore, $d'_H \leq 2n - \left\lceil \frac{2}{\ell} d_H \right\rceil + 1$ where, d'_H is the minimum Hamming distance for the binary code $\mathcal{C}$. From the Gilbert-Varshamov bound for binary code, one can obtain the bound in (2) for $d_H \geq n\ell/2$. □

Now, one can easily obtain the following proposition.

Proposition 11 *If there exists an ℓ conflict free DNA code $\mathcal{C}_{DNA}$ with all the constraints reverse, reverse-complement, GC-content such that each DNA codeword is free from secondary like structures then*

$$A_4^{cf,hf,GC,r,rc}(n, d_H, \ell) \geq \frac{2^n}{\sum_{i=0}^{2n - \left\lceil \frac{2}{\ell} d_H \right\rceil} \binom{n}{i}}, \textit{ for } d_H \geq n\ell/2.$$

In Fig. 2, the lower bound on $A_4^{cf,hf,GC,r,rc}(n, d_H, \ell)$ is plotted for n = 16 and 20.

5 Conflict free DNA codes

For various binary codes, the parameters of encoded DNA codes are listed in the Table 4. For $\mathbf{x} = ATA$ and $\mathbf{y} = CGC$, the 5 conflict free DNA code (21, 16, 6) with reverse, and GC-content constraints are obtained from [7, 4, 3] binary Hamming code is given in Table 5. The parameter of the DNA code is (21, 16, 6) and each DNA string is free from hairpin like secondary structures of stem length 5. For any positive integer $\ell = 2, 3, 4, 5$, DNA pairs $(\mathbf{x}, \mathbf{y}) \in \Sigma_{DNA}^{\ell}$ are listed in Tables 7 and 8 with various parameters. For any pair $(\mathbf{x}, \mathbf{y})$ given in Table 8, one can get ℓ conflict free DNA codes from any binary code using the Non-Homopolymer map, where the DNA code satisfies Hamming, GC-content, reverse and reverse-complement constraints. Similarly, for any pair $(\mathbf{x}, \mathbf{y})$ given in Table 7, one can get DNA codes from any binary code using the Non-Homopolymer map where, the DNA code satisfies Hamming and GC-content constraints where, each DNA codeword is free from reverse-complement sub-strings (Tables 6, 7 and 8).

5.1 Reed-Muller code:

The binary Reed-Muller codes were introduced by Reed and Muller in 1954 [25]. For two non-negative integers m and r $(r \leq m)$, the r^{th} order binary Reed-Muller code $\mathcal{R}(r, m)$ is a linear code of length 2^m, code size $2^{\sum_{i=1}^{r}\binom{m}{i}}$ and the minimum Hamming distance $d_H = 2^{m-r}$. The generator matrix of $\mathcal{R}(r, m)$ is

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

where

$$G_{m,m} = \begin{pmatrix} G_{m-1,m} \\ 1\,1\ldots1\,0 \end{pmatrix},$$

$G_{0,m}$ is the all one matrix of size 1×2^m, and $\mathbf{0}$ is a zero matrix with $2^{\sum_{i=1}^{r}\binom{m}{i}}$ - $2^{\sum_{i=1}^{r-1}\binom{m-1}{i}}$ rows and 2^{m-1} columns.

In the following theorem, the minimum distance (Definition 5) is obtained for Reed-Muller codes.

Theorem 17 *For positive integers m, r $(0 \leq r \leq m)$ and ℓ, there exists a DNA code $\mathcal{C}_{DNA}(\ell 2^m, 2^{\sum_{i=1}^{r}\binom{m}{r}}, \ell 2^{m-r-1})$ for the binary Reed-Muller code $\mathcal{R}(r, m)$.*

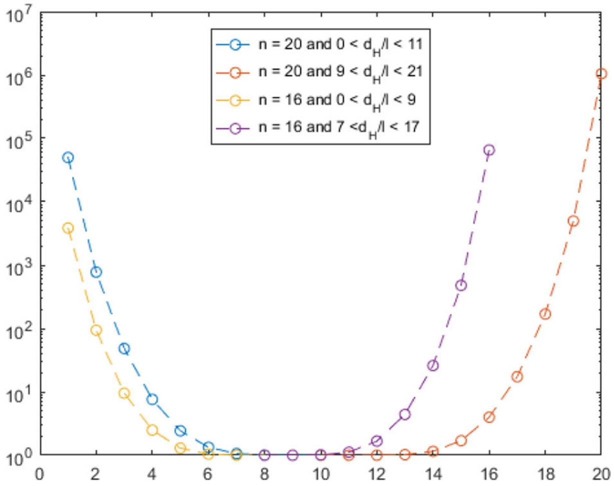

Fig. 2 For $n = 16$ and $n = 20$, graph between lower bound of $A_4^{cf,hf,GC,r,rc}(n, d_H, \ell)$ and d_H/ℓ is plotted where, the horizontal axis represents d_H/ℓ and vertical axis represents lower bound of $A_4^{cf,hf,GC,r,rc}(n, d_H, \ell)$ as calculated in Preposition 10 and Proposition 11

Table 7 Pairs $(\mathbf{x}, \mathbf{y}) \in \Sigma^{\ell}$ such that (i) GC-content sum of $\mathbf{x}$ and $\mathbf{y}$ is ℓ, and (ii) each DNA string in the set $\{(\mathbf{x}\, \mathbf{y}^*\, \mathbf{x}^*\, \mathbf{y}^*), (\mathbf{y}\, \mathbf{x}^*\, \mathbf{y}^*\, \mathbf{x}^*) : \mathbf{x}^* \in \{\mathbf{x}, \mathbf{x}^c\}$ and $\mathbf{y}^* \in \{\mathbf{y}, \mathbf{y}^c\}\}$ is free from secondary structures

ℓ	$\#(\mathbf{x}, \mathbf{y})$	$(\mathbf{x}, \mathbf{y})$
2	24	(AA, CC), (AA, GG), (CA, TC), (CA, AG), (GA, AC), (GA, TG), (CC, TT), (TC, CA), (TC, GT), (AG, CA), (AG, GT), (GG, AA), (CT, AC), (CT, TG), (GT, TC), (GT, AG), (TT, CC), (TT, GG), (AC, GA), (AC, CT), (CC, AA), (GG, TT), (TG, GA), (TG, CT)
3	32	(TAA, GCC), (TAA, GGC), (TAA, CCG), (TAA, CGG), (TTA, GCC), (TTA, GGC), (TTA, CCG), (ATT, GCC), (TTA, CGG), (GCC, TAA), (GCC, TTA), (GCC, AAT), (GCC, ATT), (GGC, TAA), (GGC, TTA), (ATT, GGC), (GGC, AAT), (GGC, ATT), (CCG, TAA), (CCG, TTA), (CCG, AAT), (CCG, ATT), (CGG, TAA), (ATT, CCG), (CGG, TTA), (CGG, AAT), (CGG, ATT), (AAT, GCC), (AAT, GGC), (AAT, CCG), (AAT, CGG), (ATT, CGG)
4	32	$(TCCA, AGGT)$, $(AGCA, TCGT)$, $(GTCA, AGTC)$, $(GTCA, TCAG)$, $(ACGA, TGCT)$, $(TGGA, ACCT)$, $(CTGA, TGAC)$, $(CTGA, ACTG)$, $(GAAC, CTTG)$, $(TGAC, CTGA)$, $(TGAC, GACT)$, $(CTAC, GATG)$, $(CATC, GTAG)$, $(AGTC, GTCA)$, $(AGTC, CAGT)$, $(GTTC, CAAG)$, $(CAAG, GTTC)$, $(TCAG, GTCA)$, $(TCAG, CAGT)$, $(GTAG, CATC)$, $(GATG, CTAC)$, $(ACTG, CTGA)$, $(ACTG, GACT)$, $(CTTG, GAAC)$, $(GACT, TGAC)$, $(GACT, ACTG)$, $(ACCT, TGGA)$, $(TGCT, ACGA)$, $(CAGT, AGTC)$, $(CAGT, TCAG)$, $(TCGT, AGCA)$, $(AGGT, TCCA)$

Proof For positive integers m (> 1) and ℓ $(\leq 2^{m-1})$, consider a binary Reed-Muller code $\mathcal{R}(r, m)$ and the corresponding encoded DNA code $f(\mathcal{R}(r, m))$ for some pair $(\mathbf{x}, \mathbf{y}) \in (\Sigma_{DNA}^{\ell})^2$. From Theorem 9, the codeword length and code size for $f(\mathcal{R}(r, m))$ will be $\ell 2^m$ and $2^{\sum_{i=1}^{r}\binom{m}{r}}$. From Theorem 15, the minimum distance $d_{NHo} \geq \ell \lceil d_H/2 \rceil = \ell 2^{m-r-1}$. For a positive integer t, we denote $\mathbf{0}_t = (0\ 0 \ldots 0)$ and $\mathbf{1}_t = (1\ 1 \ldots 1)$, each of length t. Then the binary strings $\mathbf{0}_{2^m}$ and $(\mathbf{0}_{2^m - 2^{m-r}}\ \mathbf{1}_{2^{m-r}})$ will be in $\mathcal{R}(r, m)$. Therefore $d_{NHo} \leq d_{NHo}(\mathbf{0}_{2^m}, (\mathbf{0}_{2^m-2^{m-r}}\ \mathbf{1}_{2^{m-r}})) = \ell 2^{m-r-1}$. Hence, $d_{NHo} = \ell 2^{m-r-1}$ for the binary $\mathcal{R}(r, m)$. So, from Theorem 9, the minimum Hamming distance for $f(\mathcal{R}(r, m))$ will be $d_H = \ell 2^{m-r-1}$. □

Note that by choosing appropriate seed blocks one can obtain DNA codes with various constraints. For example, if one chooses DNA seed blocks from Table 8 and constructs the Reed-Muller type code (Theorem 17) then the DNA code will satisfy reverse, reverse-complement and GC-content constraints. In addition, each DNA codeword of the Reed-Muller code is ℓ conflict free. Similarly, if one chooses DNA seed blocks from Table 7 and constructs the Reed-Muller type code (Theorem 17) then the encoded DNA code is ℓ conflict free with GC-content constraint and each codeword is free from secondary structures.

Table 8 Pairs $(\mathbf{x}, \mathbf{y}) \in \Sigma^{\ell}$ such that (i) $d_H(\mathbf{x}, \mathbf{y}) = d_H(\mathbf{x}, \mathbf{y}^{rc}) = d_H(\mathbf{x}, \mathbf{y}^{r}) = \ell$, (ii) GC-content sum of $\mathbf{x}$ and $\mathbf{y}$ is ℓ, and (iii) each DNA string in the set $\{(\mathbf{x}\ \mathbf{y}^*\ \mathbf{x}^*\ \mathbf{y}^*), (\mathbf{y}\ \mathbf{x}^*\ \mathbf{y}^*\ \mathbf{x}^*) : \mathbf{x}^* \in \{\mathbf{x}, \mathbf{x}^c\} \text{ and } \mathbf{y}^* \in \{\mathbf{y}, \mathbf{y}^c\}\}$ is ℓ conflict free

ℓ	$\#(\mathbf{x}, \mathbf{y})$	$(\mathbf{x}, \mathbf{y})$
3	8	$(ATA, CGC), (ATA, GCG), (CGC, ATA), (CGC, TAT),$ $(GCG, TAT), (TAT, CGC), (GCG, ATA), (TAT, GCG)$
4	32	$(ATCA, CGAC), (GTCA, CGAT), (ATGA, GCAG), (CTGA, GCAT),$ $(AGTA, GACG), (CGTA, GACT), (CGAC, ATCA), (TGAC, ATCG),$ $(ATGC, TCAG), (CTGC, TCAT), (AGTC, TACG), (CGTC, TACT),$ $(GACG, AGTA), (TACG, AGTC), (ATCG, TGAC), (GTCG, TGAT),$ $(GCAT, CTGA), (TCAT, CTGC), (CGAT, GTCA), (TGAT, GTCG),$ $(CAGT, GCTA), (TAGT, GCTG), (ACTA, CAGC), (GCTA, CAGT),$ $(CAGC, ACTA), (TAGC, ACTG), (GCAG, ATGA), (TCAG, ATGC),$ $(ACTG, TAGC), (GCTG, TAGT), (GACT, CGTA), (TACT, CGTC),$
5	112	$(ACGCA, CTATC), (ACGCA, GATAG), (CTGCA, GACAT),$ $(GCTCA, CTAGT), (AGCGA, CATAC), (AGCGA, GTATG),$ $(CGTGA, GTCAT), (CGTGA, GTACT), (ATCTA, GCGAC),$ $(ATCTA, GCACG), (ATCTA, CAGCG), (ATCTA, GCTCG),$ $(ATGTA, CGAGC), (ATGTA, GACGC), (ATGTA, CGTGC),$ $(ATGTA, GCTCG), (CTGTA, GCAGT), (CTGTA, GACGT),$ $(TCGAC, ATCTG), (CATAC, AGCGA), (CATAC, TCGCT),$ $(CGAGC, ATCTA), (CGAGC, ATGTA), (CGAGC, TACAT),$ $(TGAGC, ATCAG), (TGAGC, ACTAG), (AGTGC, TCATG),$ $(CGTGC, ATCTA), (CGTGC, ATGTA), (CGTGC, TACAT),$ $(TGATC, ACTCG), (CTATC, ACGCA), (CTATC, TGCGT),$ $(ATGTC, TGACG), (TACAG, ACTGC), (TACAG, ACGTC),$ $(GATAG, TGCGT), (ACTAG, TGAGC), (GCACG, TATGA),$ $(GCACG, TACAT), (GCACG, TAGAT), (GCACG, AGTAT),$ $(ACTCG, TGATC), (ACTCG, TAGTC), (GCTCG, TGATA),$ $(GCTCG, TACAT), (GCTCG, TAGAT), (GCTCG, ATAGT),$ $(GTATG, TCGCT), (AGCTG, TAGAC), (ATCTG, TCGAC),$ $(GACAT, CGTCA), (TACAT, CGAGC), (TACAT, CGTGC),$ $(TACAT, CTGCG), (TACAT, GCTCG), (CAGAT, GTCGA),$ $(TAGAT, GTCGC), (TAGAT, CGTGC), (TAGAT, GCACG),$ $(GCACT, CATGA), (GCACT, CAGTA), (GTACT, CGTGA),$ $(TCGCT, GTATG), (CGAGT, GATCA), (CGAGT, GACTA),$ $(TGCGT, CTATC), (TGCGT, GATAG), (GATCA, CGAGT),$ $(GCTCA, CTGAT), (GTCGA, CAGAT), (CATGA, GCACT),$ $(ATCTA, CGAGC), (ATCTA, CGTGC), (GTCTA, CGACT),$ $(GTCTA, CAGCT), (ATGTA, CGCAG), (ATGTA, GCACG),$ $(TAGAC, AGTCG), (TAGAC, AGCTG), (AGTAC, TCACG),$ $(CGAGC, TATCA), (CGAGC, TAGAT), (CGAGC, ACTAT),$ $(AGTGC, TACTG), (CGTGC, TCATA), (CGTGC, TAGAT),$ $(CGTGC, ATACT), (ACGTC, TACAG), (ATGTC, TGCAG),$

Table 8 (continued)

ℓ	$\#(\mathbf{x},\mathbf{y})$	$(\mathbf{x},\mathbf{y})$
		$(TGCAG, ATGTC)$, $(GATAG, ACGCA)$, $(GCACG, ATCTA)$, $(GCACG, ATGTA)$, $(TCACG, ATGAC)$, $(TCACG, AGTAC)$, $(GCTCG, ATCTA)$, $(GCTCG, ATGTA)$, $(TCATG, AGTGC)$, $(GTATG, AGCGA)$, $(ATCTG, TCAGC)$, $(GACAT, CTGCA)$, $(TACAT, GCGTC)$, $(TACAT, GCACG)$, $(CAGAT, GCTGA)$, $(TAGAT, CGAGC)$, $(TAGAT, GCTCG)$, $(TAGAT, CGCTG)$, $(CAGCT, GTCTA)$, $(TCGCT, CATAC)$, $(CTAGT, GCTCA)$, $(GACGT, CTGTA)$

6 Conclusions

We have scratched the surface of an interesting area of DNA codes that can be used in building efficient DNA data storage models.This article concentrates on two different approaches, computational and algebraic; to design DNA codes satisfying different constraints effective for practical usage. Computational approach improves the lower bounds on the size of the DNA codes in many cases from previous study under a new constraint (generalization of Homopolymers constraint) apart from the general constraints considered in the same aspect. On the contrary, the algebraic approach presents a new isometry between binary codes and DNA codes. Utilizing the recursive isometry, new classes of DNA codes has been constructed that are efficient. It is noteworthy to mention that the new codes are also free from hairpin like secondary structures. It would be an interesting future task to find bounds on DNA codes with the new constraint in mind and constructing optimal codes meeting those bounds. Extending the isometry from binary to q-ary case will also be an interesting future task.

References

1. Blawat, M., Gaedke, K., Hütter, I., Chen, X.M., Turczyk, B., Inverso, S., Pruitt, B.W., Church, G.M.: Forward error correction for DNA data storage. Procedia Comput. Sci. **80**, 1011–1022 (2016)
2. Bornholt, J., Lopez, R., Carmean, D.M., Ceze, L., Seelig, G., Strauss, K.: A DNA-based archival storage system. ACM SIGOPS Operating Syst. Rev. **50**(2), 637–649 (2016)
3. Chee, Y.M., Ling, S.: Improved lower bounds for constant GC-content DNA codes. IEEE Trans. Inf. Theory **54**(1), 391–394 (2008). https://doi.org/10.1109/TIT.2007.911167
4. Chheda, N., Gupta, M.K.: RNA As a permutation. arXiv:1403.5477v1 (2014)
5. Church, G.M., Gao, Y., Kosuri, S.: Next-generation digital information storage in DNA. Science **337**(6102), 1628–1628 (2012). https://doi.org/10.1126/science.1226355
6. Erlich, Y., Zielinski, D.: DNA Fountain enables a robust and efficient storage architecture. Science **355**(6328), 950–954 (2017). https://doi.org/10.1126/science.aaj2038
7. Gaborit, P., King, O.D.: Linear constructions for DNA codes. Theor. Comput. Sci. **334**, 99–113 (2005)
8. Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B., Birney, E.: Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature **494**(7435), 77 (2013)
9. Guenda, K., Gulliver, T.A., Solé, P.: On cyclic DNA codes. In: Proceedings IEEE International Symposium on Information Theory (ISIT), pp. 121–125. https://doi.org/10.1109/ISIT.2013.6620200 (2013)
10. Immink, K.A.S., Cai, K.: Properties and constructions of constrained codes for DNA-based data storage. arXiv:1812.06798 (2018)

11. Jacobs, A.: Data-storage for eternity (ETH Zürich, 13th of February 2015 https://www.ethz.ch/en/news-and-events/eth-news/news/2015/02/data-storage-for-eternity.html)
12. Jain, S., Hassanzadeh, F.F., Schwartz, M., Bruck, J.: Duplication-correcting codes for data storage in the DNA of living organisms. IEEE Trans. Inf. Theory **63**(8), 4996–5010 (2017). https://doi.org/10.1109/TIT.2017.2688361
13. Kari, L., Konstantinidis, S., Losseva, E., Sosík, P., Thierrin, G.: Hairpin structures in DNA words. In: DNA Computing, Pp. 158–170 (2006)
14. Kiah, H.M., Puleo, G.J., Milenkovic, O.: Codes for DNA sequence profiles. In: Proceedings IEEE International Symposium on Information Theory (ISIT), pp. 814–818. https://doi.org/10.1109/ISIT.2015.7282568 (2015)
15. Kim, Y.S., Kim, S.H.: New construction of DNA codes with constant-GC contents from binary sequences with ideal autocorrelation. In: Proceedings IEEE International Symposium on Information Theory (ISIT), pp. 1569–1573. https://doi.org/10.1109/ISIT.2011.6033808 (2011)
16. Kovačević, M., Tan, V.Y.F.: Asymptotically optimal codes correcting fixed-length duplication errors in DNA storage systems. IEEE Commun. Lett. **22**(11), 2194–2197 (2018). https://doi.org/10.1109/LCOMM.2018.2868666
17. Limbachiya, D., Benerjee, K.G., Rao, B., Gupta, M.K.: On DNA codes using the ring $\mathbb{Z}_4 + w\mathbb{Z}_4$. In: Proceedings IEEE International Symposium on Information Theory (ISIT), pp. 2401–2405. https://doi.org/10.1109/ISIT.2018.8437313 (2018)
18. Limbachiya, D., Gupta, M.K.: Natural Data Storage: A Review on sending Information from now to then via Nature. arXiv:1505.04890 (2015)
19. Limbachiya, D., Gupta, M.K., Aggarwal, V.: Family of constrained codes for archival DNA data storage. IEEE Commun. Lett. **22**(10), 1972–1975 (2018). https://doi.org/10.1109/LCOMM.2018.2861867
20. Limbachiya, D., Rao, B., Gupta, M.K.: The Art of DNA Strings: Sixteen Years of DNA Coding Theory. arXiv:1607.00266 (2016)
21. Loman, N., Misra, R., Dallman, T., Constantinidou, C., Gharbia, S., Wain, J., Pallen, M.: Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. **30**(6), 434–439 (2012)
22. Lothaire, M. Combinatorics on Words, 2nd edn. Cambridge Mathematical Library. Cambridge University Press, Cambridge (1997). https://doi.org/10.1017/CBO9780511566097
23. Marathe, A., Condon, A.E., Corn, R.M.: On combinatorial DNA word design. J. Comput. Biol. **8**(3), 201–219 (2001). https://doi.org/10.1089/10665270152530818
24. Milenkovic, O., Kashyap, N.: DNA Codes that avoid secondary structures. In: Proceedings IEEE International Symposium on Information Theory (ISIT), pp. 288–292. https://doi.org/10.1109/ISIT.2005.1523340 (2005)
25. Muller, D.E.: Application of boolean algebra to switching circuit design and to error detection. Transactions of the I. R. E. Professional Group on Electronic Computers **EC-3**(3), 6–12 (1954). https://doi.org/10.1109/IREPGELC.1954.6499441
26. Myers, P., Sebaihia, M., Cerdeño-tárraga Bentley, S., Crossman, L., Parkhill, J.: Tandem repeats and morphological variation. Nature (2007)
27. Nelms, B.L., Labosky, P.A.: A predicted hairpin cluster correlates with barriers to PCR. sequencing and possibly BAC recombineering. Scientific Reports **1**, 106 (2011)
28. Ridge, P., Carroll, H., Sneddon, D., Clement, M., Snell, Q.: Large grain size stochastic optimization alignment. In: Proceedings IEEE Symposium on BioInformatics and BioEngineering (BIBE), pp. 127–134. https://doi.org/10.1109/BIBE.2006.253325 (2006)
29. Rykov, V.V., Macula, A.J., Torney, D.C., White, P.S.: DNA Sequences and quaternary cyclic codes. In: Proceedings IEEE International Symposium on Information Theory (ISIT), pp. 248–248. https://doi.org/10.1109/ISIT.2001.936111 (2001)
30. Smith, D.H., Aboluion, N., Montemanni, R., Perkins, S.: Linear and nonlinear constructions of DNA codes with Hamming distance d and constant GC-content. Discret. Math. **311**(13), 1207–1219 (2011)
31. Song, W., Cai, K., Zhang, M., Yuen, C.: Codes with run-length and GC-content constraints for DNA-based data storage. IEEE Commun. Lett. **22**(10), 2004–2007 (2018). https://doi.org/10.1109/LCOMM.2018.2866566
32. Thomson, N., Sebaihia, M., Cerdeño-tárraga Bentley, S., Crossman, L., Parkhill, J.: The value of comparison. Nat. Rev. Microbiology **1**(11), 11–12 (2003)
33. Tulpan, D., Smith, D.H., Montemanni, R.: Thermodynamic post-processing versus GC-content pre-processing for DNA codes satisfying the hamming distance and reverse-complement constraints. IEEE/ACM Trans. Comput. Biol. Bioinform. **11**(2), 441–452 (2014). https://doi.org/10.1109/TCBB.2014.2299815
34. Tulpan, D.C., Hoos, H.H., Condon, A.E.: Stochastic local search algorithms for DNA word design. In: DNA Computing, pp. 229–241 (2003)

35. Yakovchuk, P., Protozanova, E., Frank-Kamenetskii, M.D.: Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nuclice Acis Res. **34**(2), 564–574 (2006). https://doi.org/10.1093/nar/gkj454
36. Yazdi, S.H.T., Yuan, Y., Ma, J., Zhao, H., Milenkovic, O.: A rewritable, random-access DNA-based storage system. Scientific Reports **5**, 14138 (2015)
37. Zhu, X., Sun, C., Liu, W., Wu, W.: Research on the counting problem based on linear constructions for DNA coding. In: Proceedings Computational Intelligence and Bioinformatics, pp. 294–302 (2006)
38. Zuker, M.: Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31**(13), 3406–3415 (2003)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.