On Cyclic DNA Codes

Kenza Guenda and T. Aaron Gulliver

May 2, 2014

Abstract

This paper considers cyclic DNA codes of arbitrary length over the ring $R = \mathbb{F}_2[u]/u^4 - 1$. A mapping is given between the elements of $R$ and the alphabet \{A, C, G, T\} which allows the additive stem distance to be extended to this ring. Cyclic codes over $R$ are designed such that their images under the mapping are also cyclic or quasi-cyclic of index 2. The additive distance and hybridization energy are functions of the neighborhood energy.

1 Introduction

Deoxyribonucleic acid (DNA) contains the genetic program for the biological development of life. DNA is formed by strands linked together and twisted in the shape of a double helix. Each strand is a sequence of four possible nucleotides, two purines: adenine (A) and guanine (G), and two pyrimidines: thymine (T) and cytosine (C). The ends of a DNA strand are chemically polar with 5′ and 3′ ends, which implies that the strands are oriented. Hybridization, also known as base pairing, occurs when two strands bind together, forming a double strand of DNA. The strands are linked following the Watson-Crick model, so that every A is linked with a T, and every C with a G, and vice versa. We denote the complement of $x$ as $\hat{x}$, i.e., $\hat{A} = T, \hat{T} = A, \hat{G} = C$ and $\hat{C} = G$. DNA strand pairing is done
in the opposite direction and the reverse order. For instance, the Watson-
Crick complementary (WCC) strand of $3' - ACT T A G A - 5'$ is the strand
$5' - T C T A A G T - 3'$. Nucleotide pairing is based on hydrogen bonds, with
a pair $A - T$ forming two bonds, a pair $G - C$ forming three bonds, and any
other pair is called a mismatch because it does not form a bond.

The combinatorial properties of DNA sequences can be used to tackle
computationally difficult problems. For example, Adleman [3] solved an in-
stance of a hard (NP-complete) computational problem, namely the directed
traveling salesman problem on a graph with seven nodes. Adleman et al. [4]
used the WCC approach to break the data encryption standard (DES). In
addition, Lipton [20] used DNA strands to solve the satisfiability (SAT) prob-
lem. Further, Ouyang et al. [27] presented a DNA solution to the maximum
clique problem. Since there are $4^n$ possibly single DNA strands of length $n$
which can quickly and cheaply be synthesized, Mansuripur et al. [22] showed
that DNA codewords can be used for ultra high density data storage. Other
applications exploit DNA hybridization [29].

Software such as AMBER or CHARMM exist which can provide an ac-
curate representation of the DNA molecule. However, these methods are
computational demanding and have a time scale on the order of $\mu s$. This
creates difficulties as many biological and other processes have time scales
on the order of $ms$. Further, these packages do not allow study of the DNA
hybridization of strands (duplex formation from single strands). This is a
significant problem, as hybridization can be used as a gate in a DNA com-
puter. To allow parallel operations on DNA sequences, a high hybridization
energy is required. This energy depends in a rather complex way on the
number of hydrogen bonds and their arrangement in the duplex. A duplex
formed by a single strand with high $GC$ content and its reverse complemen
t has greater stability since this pair has a high number of hydrogen bonds.
Note that in this case there are no mismatches in the duplex. Hence the
importance of designing groups of DNA words, called a DNA code, which
satisfy the reverse complement constraint.

Breslauer et al. [8] introduced the nearest-neighbor similarity model in
order to estimate the hybridization energy of a duplex. In this model, the
energy is a sum taken over pairs of positions rather than single positions. For example, the energy of $3' - CATG - 5'$ is equal to $e(CA/GT) + e(AT/TA) + e(TG/AC)$, where $e(\ )$ is the neighborhood energy of the pairs formed by the nucleotides and their WC complements, which are called stacked pairs. This energy has been determined by experimental methods and a comprehensive survey of these results is given in [28]. This model can be used in the ideal case, i.e., when a single strand hybridizes with its WC complement, which is not always the case.

Secondary structure occurs when a strand folds back onto itself forming a double strand. Milenkovic and Kashyap [25] argued that when designing a DNA code, a cyclic constraint should be added to reduce the probability of secondary structure. Secondary structure causes codewords to become computationally inactive. This defeats the read-back mechanism in a DNA storage system by as much as 30% as reported by Mansuripur et al. [22]. Milenkovic and Kashyap [25] used the Nussinov-Jacobson algorithm [26] to prove that the presence of a cyclic structure reduces the complexity of testing DNA codes for secondary structure.

There have been numerous results on the design of DNA codes [1, 2, 6, 16, 18]. The problem of hybridization energy has also been studied extensively [5, 11, 14, 30]. More recently, D’Yachkov et al. [15] modeled the hybridization energy for DNA strand as an additive stem similarity using the neighborhood energy of pairs of nucleotides. They also introduced the additive stem distance. Bahattin and Siap [6] constructed DNA codes as cyclic reversible-complement codes of odd length over the ring

$$R = \mathbb{F}_2[u]/(u^4 - 1) = \{a + bu + cu^2 + du^3 \mid a, b, c, d \in \mathbb{F}_2, u^4 = 1\}.$$  

They also studied the problem of the Hamming distance.

In this paper we construct cyclic DNA codes of arbitrary length over the ring $R = \mathbb{F}_2[u]/u^4 - 1$. This is a finite chain ring with 16 elements. A mapping is given between the elements of $R$ and the alphabet $\{A, C, G, T\}$ which allows the notion of additive stem distance to be extended to this ring. Cyclic codes are obtained over $R$ which are reversible-complement and have images under the mapping which are also cyclic or quasi-cyclic of
index 2. They also satisfy the WCC condition, and the additive distances and hybridization energies can be determined. Note that one can also find a one-to-one to map between the elements \{A, C, G, T\}^2 and the field of cardinality 16, but codes over a ring are more suitable. In particular, codes over rings can contain more codewords than similar codes over fields, and they provide more flexibility in constructing codes. Moreover, there exists more cyclic codes over rings than over fields. The structure of repeated cyclic codes over finite chain rings is in general not known. We use the fact that \( \mathbb{F}_2 \subset R \) and they have the same characteristic to find the structure of these codes. Note that the results given hold for any finite chain ring with cardinality 16 and characteristic 2. For example, \( \mathbb{F}_2 + u\mathbb{F}_2 + u^2\mathbb{F}_2 + u^3\mathbb{F}_2 \) with \( u^4 = 0 \) is such a ring. Typically the Hamming distance or deletion distance are used in designing DNA codes. However, these metrics do not capture the thermodynamic properties and combinatorial structure of DNA. We consider the additive stem distance and adapt it for use with our DNA codes. Another reason for using the ring \( R \) is that the codes can be mapped to DNA codes of length \( 2n \) which contain a subcode with large GC-content and thus has a high hybridization energy.

The next section presents some basic facts and preliminaries. Section 3 introduces the additive stem-similarity model. Then cyclic DNA codes are investigated in Section 4. In particular, the structure of cyclic codes of even length over the ring \( R \) is determined. Our DNA codes are presented and the stem-similarity distance is extended to these codes.

## 2 Preliminaries

The ring considered here is

\[
R = \mathbb{F}_2[u]/(u^4 - 1) = \{a + bu + cu^2 + du^3 \mid a, b, c, d \in \mathbb{F}_2, u^4 = 1\},
\]

which is a commutative ring with 16 elements. It is a principal local ideal with maximal ideal \( \langle u + 1 \rangle \). The ideals satisfy

\[
\langle 0 \rangle = \langle (u + 1)^4 \rangle \subsetneq \langle (u + 1)^3 \rangle \subsetneq \langle (u + 1)^2 \rangle \subsetneq \langle (u + 1) \rangle \subsetneq R.
\] (1)
Table 1: Correspondence between the nucleotide base pairs and the elements of $R$.

|   | GG    | AT  | 1 + u | GT  | 1 | CT  | 1 + u + u^2 |
|---|-------|-----|-------|-----|---|-----|-------------|
| CC | 1 + u + u^2 + u^3 | TA  | u^2 + u^3 | TG  | u^2 | TC  | 1 + u^2 + u^3 |
| GC | 1 + u^2 | AA  | u + u^2 | AC  | 1 + u + u^3 | AG  | u            |
| CG | u + u^3 | TT  | 1 + u^3 | CA  | u + u^2 + u^3 | GA  | u^3          |

The field $\mathbb{F}_2$ is a subring of $R$, a fact which will be used later.

A map $\phi$ which is a one-to-one correspondence between the elements of $R$ and the DNA nucleotide base pairs $\{A, T, C, G\}^2$ is given in Table 1. A simple verification gives that for all $x \in R$, we have

$$x + \hat{x} = u^3 + u^2 + u + 1. \tag{2}$$

In addition, multiplying an element $x$ of $R$ by $u^2$ reverses the DNA pair corresponding to $x$. Further, multiplying any $x \in R$ by $u^2$ reverses the corresponding pair in $\{A, G, C, T\}^2$. Note that other mappings can be defined between $R$ and the nucleotide pairs [6]. The mapping $\phi$ was chosen because it results in a subcode over the alphabet $\{GC, CC, GG, CG\}$ which will have a high hybridization energy.

Since $R^n$ is an $R$ module, a linear code over $R$ of length $n$ is a submodule $C$ of $R^n$. Now let $A$ any alphabet. Then a code $C$ over $A$ is said to be cyclic if it is invariant under a cyclic shift, i.e., $(x_{n-1}, x_0, \ldots, x_{n-2}) \in C$ provided the codeword $(x_0, x_1, \ldots, x_{n-2}, x_{n-1})$ is in $C$. A code is called quasi-cyclic of index $l$ if for any $(x_0, x_1, \ldots, x_{n-2}, x_{n-1}) \in C$ we have $(x_{n-l}, x_{n-l+1}, \ldots, x_0, x_1, x_{n-l-1}) \in C$. Note that these definitions hold regardless of whether the code is linear.

The structure of linear cyclic codes of length $n$ over $R$ when $n$ is odd has been examined in [9][10], but the general case has not yet been investigated.

Let $x = x_0 x_1 \ldots x_{n-1}$ be a vector in $R^n$. The reverse of $x$ is defined to be $x^r = x_{n-1} x_{n-2} \ldots x_1 x_0$, the complement of $x$ is $x^c = \hat{x}_0 \hat{x}_1 \ldots \hat{x}_{n-1}$, and the reverse complement (also called the Watson-Crick complement) is
$x^{rc} = x_{n-1} \hat{x}_{n-2} \ldots \hat{x}_1 \hat{x}_0$. A code $C$ is said to be **reverse complement** if for any $x \in C$ we have $x^{rc} \in C$.

**Definition 2.1** For $\mathcal{A} = \{A, G, C, T\}$, a code $C$ of length $n$ over the alphabet $\mathcal{A}$ is called reversible if the WCC of each codeword $a \in \mathcal{A}^n$ is also in $C$.

### 3 Additive Stem Similarity Distance

In this section, the additive stem similarity introduced by D’yachkov and Voronina [15] is presented. The DNA hybridization energy for strands $x$ and $y$ is an important measure of the stability of the duplex, as it is related to the melting temperature of the duplex. This is the temperature required to melt a duplex. The hybridization energy of a duplex can be modeled as a function of the so-called neighborhood energy of the nucleotides. For a pair $a, b \in \mathcal{A} = \{A, C, G, T\}$, the neighborhood energy is given by

$$w(a, b) = \Delta G(a, b) = \Delta H(a, b) - T \Delta S(a, b),$$

where $\Delta H(a, b)$ and $\Delta S(a, b)$ are the temperature-independent enthalpy and entropy, respectively. The pairs $(a, b) \in \mathcal{A}$ are also called stacked pairs. For example, these quantities as well as $\Delta G(a, b)$ for a temperature of $310^\circ$ are given in Table 2. For $x = x_1, \ldots, x_n \in \mathcal{A}^n$ and $y = y_1, \ldots, y_n \in \mathcal{A}^n$, define

$$S_w(x, y) = \sum_{i=1}^{n-1} s^w_i(x, y),$$

where

$$s^w_i(x, y) = \begin{cases} w(a, b) & \text{if } x_i = y_i = a, x_{i+1} = y_{i+1} = b, \\ 0 & \text{otherwise.} \end{cases}$$

and $w(a, b)$ is the neighborhood energy of the pair $(a, b) \in \mathcal{A}^2$. The quantity $S_w(x, y)$ is called the additive stem similarity between $x$ and $y$, and it satisfies the following properties

$$S_w(x, y) = S_w(y, x) \leq S_w(x, x).$$
Table 2: Nearest Neighbor Thermodynamic Values for Stacked Pairs

| Stacked pair         | $\Delta H$ kcal/mol | $\Delta S$ kcal/mol | $\Delta G_{310^\circ}$ kcal/mol | n   |
|----------------------|----------------------|----------------------|---------------------------------|-----|
| $5'$ → $3'$/$3'$ → $5'$ |                      |                      |                                 |     |
| $AA/TT = TT/AA$      | −7.9                 | −22.2                | −1.02                           |     |
| $AC/TG = GT/CA$      | −8.4                 | −22.4                | −1.46                           |     |
| $AG/TG = CT/GA$      | −7.8                 | −21.0                | −1.29                           |     |
| $AT/TA$              | −7.2                 | −20.4                | −0.88                           |     |
| $CA/GT = TG/AC$      | −8.5                 | −22.7                | −1.46                           |     |
| $CC/GG = GG/GC$      | −8.0                 | −19.9                | −1.83                           |     |
| $CG/GC$              | −10.6                | −27.2                | −2.17                           |     |
| $GA/CT = TC/AG$      | −8.2                 | −22.2                | −1.32                           |     |
| $GC/CG$              | −9.8                 | −24.4                | −2.24                           |     |
| $TA/AT$              | −7.2                 | −21.3                | −0.60                           |     |

The hybridization energy between $x$ and $y$ is

$$E(x, y) = S_w(x, y^{rc}).$$

**Definition 3.1** Let $x$ and $y$ in $A^n$. Then the real number

$$D(x, y) = S_w(x, x) - S_w(x, y)$$

is called the additive stem distance between $x$ and $y$ in $A^n$.

It is clear that $D(x, x) = 0$, but in general it is not symmetric and does not satisfy the triangle inequality.

4 Cyclic DNA Codes

Let $\mathcal{C}$ be a linear code over $R$. $A = \{A, C, G, T\}$ and $D(., .)$ be the additive stem distance given in Definition 3.1. Since the map $\phi$ defined in Table 1
is one-to-one, the additive stem distance can be extended to the ring $R$ as follows. For $x, y \in R$, define the additive stem distance over $R$ as

$$D(x, y) = D(\phi(x), \phi(y)).$$  \hspace{1cm} (3)$$

Let $D = \min_{x \neq y} D(x, y)$ for $x, y \in C$. A cyclic DNA code over $R$ is then defined as follows.

**Definition 4.1** A cyclic code $C$ over $R$ is called an $[n, d]$ cyclic DNA code if it satisfies the following:

- $C$ is a cyclic code, i.e., $C$ is an ideal in $R_n = R[x]/(x^n - 1)$;
- for any codeword $x \in C$, $x \neq x^{rc}$ and $x^{rc} \in C$; and
- $D(x, y) \geq d, \forall x, y \in C$.

Let $C$ be an $[n, d]$ cyclic DNA code over $R$. Then if $s = \max\{S_w(\phi(x), \phi(x)), x \in C\}$, from (3) and the definition of the additive stem distance we obtain that

$$S_w((\phi(x), \phi(y)) \leq S_w((\phi(x), \phi(x)) - D((\phi(x), \phi(y)), \forall x, y \in C.$$  

Therefore $S_w(\phi(x), \phi(y)) \leq s - d$ for all $x, y \in C$, and thus in our context a cyclic DNA code over $R$ is a cyclic reverse-complement code such that

$$E(\phi(x), \phi(y)) \leq s - d, \forall x, y \in C. \hspace{1cm} (4)$$

**Definition 4.2** A code $C$ over an alphabet $A$ is called an $(n, d)$ DNA code if it is a block code of length $n$ such that $D(x, y) \geq d$ for all $x, y \in C$.

### 4.1 Cyclic Codes over $R$ of Arbitrary Length

The purpose of this section is to determine the structure of cyclic codes over the ring $R$. Only codes of even length are considered as the structure of cyclic codes over $R$ of odd length has previously been examined [9, 10]. In the case $n$ odd it has been proven that the cyclic codes over $R$ are in fact principal ideals. This is not true for the case $n$ even.

We begin by providing some results for codes of odd length.
Lemma 4.3  A cyclic code of odd length $n$ over $R$ is an ideal defined as 

$$C = \langle f_0 | (u+1)f_1 | (u+1)^2 f_2 | (u+1)^3 f_3 \rangle$$

(5) such that $f_3 | f_2 | f_1 | f_0 | x^n - 1$.

Note that there exists a canonical surjective ring morphism $(-)$ given by 

$$(-) : R[x] \longrightarrow \mathbb{F}_2[x]$$

$$f \mapsto \overline{f} = f \mod u + 1$$

(6)

Definition 4.4  A polynomial $f$ in $R[x]$ is called regular if $f \neq 0$. $f$ is called primary if the ideal $\langle f \rangle$ is primary, and $f$ is called basic irreducible if $\overline{f}$ is irreducible in $\mathbb{F}_2[x]$. Two polynomials $f$ and $g$ in $R[x]$ are called coprime if $R[x] = \langle f \rangle + \langle g \rangle$.

Lemma 4.5  (\cite{24}, Theorem XIII. 11) Let $f$ be a regular polynomial in $R[x]$. Then $f = \alpha g_1 \ldots g_r$, where $\alpha$ is a unit and $g_1, \ldots, g_r$ are regular primary coprime polynomials. Moreover, $g_1, \ldots, g_r$ are unique in the sense that if $f = \alpha g_1 \ldots g_r = \beta h_1 \ldots h_s$, where $\alpha, \beta$ are units and $g_i$ and $h_i$ are regular primary coprime polynomials, then $r = s$, and after renumbering $\langle g_i \rangle = \langle h_i \rangle$, $1 \leq i \leq n$.

Lemma 4.6  If $f(x) \in R[x]$ is a basic irreducible polynomial, then $f(x)$ is a primary polynomial.

Proof. Assume that $f(x)$ is basic irreducible and $g(x)h(x) \in \langle f(x) \rangle$. Then $\overline{f}(x)$ is irreducible in $K[x]$, so that $(\overline{f}(x), \overline{g}(x)) = 1$ or $\overline{f}(x)$. If $(\overline{f}(x), \overline{g}(x)) = 1$ then $f$ and $g$ are also coprime, and there exist $f_1$ and $g_1$ in $R[x]$ such that $1 = f(x)f_1(x) + g(x)g_1(x)$. Hence $h(x) = f(x)h(x)f_1(x) + g(x)h(x)g_1(x)$. Since $g(x)h(x) \in \langle f(x) \rangle$, it follows that $h(x) \in \langle f(x) \rangle$. If $(\overline{f}(x), \overline{g}(x)) = \overline{f}(x)$, then there exist $f_1(x), g_1(x) \in R[x]$ such that $g(x) = f(x)f_1 + (u+1)^i g_1(x)$ for some positive integer $i < 4$. Then for $k > i$, we have $g^k \in \langle f(x) \rangle$, and thus $f(x)$ is a primary polynomial. \qed
Remark 4.7 Let $m$ be an odd integer. Then from [19] the polynomial $x^m - 1$ factors uniquely as a product of monic basic irreducible pairwise coprime polynomials over $R$, and there is a one-to-one correspondence between the set of irreducible divisors in $\mathbb{F}_2$. Since $\mathbb{F}_2$ is a subring of $R$ and the decomposition of $x^m - 1$ is unique in $R$, the polynomials $f_i$ are in $\mathbb{F}_2$.

Proposition 4.8 If $n = m2^s$ such that $m$ is an odd integer, then $x^n - 1$ has a unique decomposition over $R$ given by

$$x^n - 1 = g_1^{2^s} \ldots g_l^{2^s},$$

(7)

where the $g_i$ are irreducible polynomials coprime in $\mathbb{F}_2[x]$ which are divisors of $x^m - 1$.

Proof. For any integer $s \geq 0$ and odd $m \geq 0$. We have that $(x^m - 1)^{2^s} = x^{m2^s} - 1$ because $2 | (2^s)$ for $1 \leq i \leq 2^s$. From Remark 4.7, $x^m - 1$ has a unique decomposition into irreducible polynomials over $\mathbb{F}_2$ as follows: $x^m - 1 = g_1 \ldots g_l$. We need to prove that $x^n - 1 = g_1^{2^s} \ldots g_l^{2^s}$ is unique. Assume that $x^n - 1 = f_1^{a_1} \ldots f_r^{a_r}$ is a decomposition into powers of basic irreducible polynomials. From Lemma 4.6 we have that the basic irreducible polynomials are primary, hence the power of a basic irreducible polynomial is also a primary polynomial. Then from Lemma 4.5, the decomposition (7) is unique.

Proposition 4.9 With the previous notation, the primary ideals of $\mathcal{R}$ are $\langle 0 \rangle$, $\langle 1 \rangle$, $\langle g^j_1 \rangle$, $\langle g^j_i, (u + 1)^t \rangle$, with $1 \leq j \leq 2^s$, $1 \leq t \leq 3$ and $1 \leq i \leq l$.

Proof. Let $\mu : R[x] \longrightarrow \frac{\mathbb{F}_2[x]}{(x^n - 1)}$ be the canonical homomorphism. By Lemma 4.8 we have that the factorization of $x^n - 1 = g_1^{2^s} \ldots g_l^{2^s}$ over $R[x]$ is the same as that over $\mathbb{F}_2[x]$ and is unique. This gives that the kernel of $\mu$ is the ideal $\langle x^n - 1, u \rangle$. Hence from [32, Theorem 3.9.14], the primary ideals of $\mathcal{R}$ are the preimages of the primary ideals of $\mathbb{F}_2[x]/x^n - 1$. It is well known [17, Theorem 3.10] that the primary ideals of this last ring are the ideals $\langle g^j_i \rangle$, $1 \leq j \leq 2^s$ and $1 \leq i \leq l$. Hence the primary ideals of $\mathcal{R}$ are $\langle g^j_i, (u + 1)^t \rangle$. $\square$
Theorem 4.10  Let $n = m2^s$ such that $m$ is an odd integer. Then the cyclic codes of length $2^sm$ over $R$ are the ideals generated by $(f_0|u+1) f_1|(u+1)^2 f_2|(u+1)^3 f_3)$, where $f_3|f_2|f_1|f_0|x^n - 1$.

Proof. Let $C$ be a cyclic code in $R[x]$ so that $C$ is an ideal of $R$. Since $R$ is Noetherian, from the Lasker-Noether decomposition Theorem [32, p. 209], any ideal in $R$ has a representation as a product of primary ideals. From Proposition 4.9, we have that the primary ideals of $R$ are $(g_i^l,(u+1)^l)$, where $x^n - 1 = \prod_{k=1}^r g_i^{2^k}$. Hence an ideal $I$ of $R$ is of the form

\[ I = \prod_{l=1}^r (g_i^l,(u+1)^l). \] (8)

Expanding the product in (8), each ideal in $R$ is generated by

\[ (f_0|(u+1) f_1|(u+1)^2 f_2|(u+1)^3 f_3), \]

where $f_3|f_2|f_1|f_0|x^n - 1$. \qed

4.2 The Reverse-Complement Constraint

In this section, the reverse-complement constraint is examined for cyclic codes of arbitrary length $n$. Denote $(x^n - 1)/(x - 1) = \mathbb{I}(x)$. The following lemma will be used later.

Lemma 4.11 ( [2]) Let $f(x)$ and $g(x)$ be polynomials in $R[x]$ with $\deg f(x) \geq \deg f(x)$. Then the following holds:

(i) $[f(x)g(x)]^* = f(x)^*g(x)^*$;

(ii) $[f(x) + g(x)]^* = f(x)^* + x^{\deg f - \deg g} g(x)^*$.

Theorem 4.12 Let $C$ be a reverse-complement cyclic code over $R$. Then the following holds:

(i) $C$ contains the codeword $(1 + u + u^2 + u^3)\mathbb{I}(x)$. 

11
(ii) $\mathcal{C} = \langle f_0 | (u + 1) f_1 | (u + 1)^2 f_2 | (u + 1)^3 f_3 \rangle$ with all $f_i$ self-reciprocal.

**Proof.** Part (i) is from [6]. Part (ii) in the case $n$ odd was proven in [6, Theorem 4.3]. Since by Theorem 4.10 the codes of even length are generated by $\mathcal{C} = \langle f_0 | (u + 1) f_1 | (u + 1)^2 f_2 | (u + 1)^3 f_3 \rangle$, the argument in [6] for $n$ odd also holds for the case $n$ even. □

The proof of the following Theorem is the same as that of [6, Theorem 4.4] for odd length.

**Theorem 4.13** Let $\mathcal{C}$ be a cyclic code over $R$ of length $n$. Suppose $(1 + u + u^2 + u^3) I(x) \in \mathcal{C}$. Then if $\mathcal{C} = \langle f_0 | (u + 1) f_1 | (u + 1)^2 f_2 | (u + 1)^3 f_3 \rangle$, where the $f_i$ are self-reciprocal, then $\mathcal{C}$ is a reverse-complement code.

**Corollary 4.14** Let $\mathcal{C}$ be a cyclic code of length $n = 2^s m$, $s \geq 0$. Then if $(1 + u + u^2 + u^3) I(x) \in \mathcal{C}$ and if there exists an $i$ such that

$$2^i \equiv -1 \mod m,$$

then $\mathcal{C}$ is a reverse-complement code.

**Proof.** Let $\mathcal{C} = \langle f_0 | (u + 1) f_1 | (u + 1)^2 f_2 | (u + 1)^3 f_3 \rangle$ be a cyclic code of length $n$. The polynomials $f_i$ are divisors of $x^n - 1$ in $\mathbb{F}_2$. The decomposition into the product of minimal polynomials is given by $x^n - 1 = \prod M_i(x)$. Each $M_i$ corresponds to a cyclotomic class $Cl(i)$. Equation (9) gives that $Cl(1)$ is reversible and hence all the cyclotomic classes are reversible. Thus each minimal polynomial is self-reciprocal, and from Lemma 4.11 the polynomials $f_i$ are self-reciprocal. Then from Theorem 4.13 $\mathcal{C}$ is a reverse-complement code. □

**Example 4.15** Let $n = 6$. Then the cyclic code over $R$ with generator polynomial $(1 + u + u^2 + u^3)(x^2 + x + 1)$ is a cyclic reversible code over $R$.

**Corollary 4.16** Let $\mathcal{C}$ be an $[n, d]$ cyclic DNA code over $R$. Then $\phi(\mathcal{C})$ is a $[2n, d]$ quasi-cyclic DNA code of index 2 over the alphabet $\{A, G, C, T\}$. 12
Proof. Let $C$ be a cyclic DNA code of length $n$ over $R$. Hence $\phi(C)$ is a set of length $2n$ over the alphabet $A$ which is quasi-cyclic of index 2. Since $C$ is a reverse-complement code, then $u^2x^{rc} \in C$, and $\phi(u^2x^{rc})$ is the WCC of $\phi(x)$. \hfill \Box

Definition 4.17 For a code $C$ over $R$, define the subcode $C_{1+u^2}$ consisting of all codewords in $C$ that are a multiple of $(1 + u^2)$.

Lemma 4.18 With the previous definition we have

$$\phi((1 + u^2)R) = \{GG, CC, CG, GC\}.$$  

Further, if $C = \langle f_0\vert (u + 1)f_1\vert (u + 1)^2f_2\vert (u + 1)^3f_3 \rangle$ is a cyclic code of length $n$ over $R$, then

$C_{1+u^2} = \langle (1 + u^2)f_3(x) \rangle$.

Proof. The first part of the lemma is a simple verification. For the second part, assume $C = \langle f_0\vert (u + 1)f_1\vert (u + 1)^2f_2\vert (u + 1)^3f_3 \rangle$. Since $f_3|f_2$ then we have $\langle (1 + u^2)f_3 \rangle \subset C_{1+u^2}$.

Now let $c(x) \in C$ so that $c(x) = k_0(x)f_0(x) + (u + 1)k_1(x)f_1(x) + (u + 1)^2k_2(x)f_2(x) + (u + 1)^3k_3(x)f_3(x)$ for $k_i(x) \in F_2[x]$. If $c(x)$ is a multiple of $1 + u^2$, then we have $x^n - 1|k_0(x)f_0(x)$ and $x^n - 1|k_0(x)f_1(x)$ and hence $c(x) = (1 + u^2)((k_2(x)f_2(x) + (1 + u)(k_3(x)f_3(x))$. Since $f_3(x)|f_2(x)|f_1(x)$, $C_{1+u^2} \subset \langle f_3(x) \rangle$, and therefore $C_{1+u^2} = \langle (1 + u^2)f_3 \rangle$. \hfill \Box

Let $d_{1+u^2} = \min\{D(x, y), x, y \in C_{1+u^2}\}$. Then the following holds.

Theorem 4.19 Let $C = \langle f_0\vert (u + 1)f_1\vert (u + 1)^2f_2\vert (u + 1)^3f_3 \rangle$ be an $[n, d]$ cyclic DNA code over $R$. Then $\phi(C_{1+u^2})$ is a cyclic DNA code of length $n$ over the alphabet $\{GG, CC, GC, CG\}$ such that $d_{1+u^2} \geq d$.

Proof. Since $C_{1+u^2} \subset C$, it is obvious that $d_{1+u^2} \geq d$. From Theorem 4.16 we have that the image of the cyclic DNA code $C$ obtained via $\phi$ is a quasi-cyclic code of length $2n$ over the alphabet $\{A, G, C, T\}$. From Lemma 4.18 we have $\phi((1 + u^2)R) = \{GG, CC, CG, GC\}$.
we have that $\phi((1+u^2)R) = \{GG, CC, GC, GC\}$ and $C_{1+u^2} = \langle(1+u^2)f_3(x)\rangle$. This gives the result. □

Remark 4.20 Theorem 4.19 is useful as it results in cyclic subcodes with large GC-content. Since from Table 2 the stacked pair corresponding to \{GG, CC, GC, CG\} has the largest neighborhood energy, these subcodes high hybridization energy.

Acknowledgements

The authors would like to thank Anne Condon for drawing their attention to the problem of the nearest neighbor energy model.

References

[1] T. Abualrub, A. Ghrayeb and X. Nian Zeng, Construction of cyclic codes over GF(4) for DNA computing, J. Franklin Institute, 343(4-5), 448–457, 2006.

[2] T. Abualrub and I. Siap, Cyclic codes over the rings $\mathbb{Z}_2 + u\mathbb{Z}_2$ and $\mathbb{Z}_2 + u\mathbb{Z}_2 + u^2\mathbb{Z}_2$, Des., Codes, Cryptog. 42(3), 273–287, 2007.

[3] L. M. Adleman, Molecular computation of solutions to combinatorial problems, Science, (266), 1021–1024, Nov. 1994.

[4] L. M. Adleman, P. W. K. Rothmund, S. Roweis, and E. Winfree, On applying molecular computation to the Data Encryption Standard, Proc. Int. DIMACS Meeting on DNA Computers, 1996.

[5] M. A. Bishop, A. G. D’Yachkov, A. J. Macula, T. Renz, and V. Rykov, Free energy gap and statistical theormodynamic fidelity of DNA codes, J. Comp. Biology, 14(8), 1088–1104, Oct. 2007.

[6] B. Yildiz and I. Siap, Cyclic DNA codes over the ring $F_2[u]/(u^4−1)$ and applications to DNA codes, Com. Math App. 63(7), 1169–1176, Apr. 2012.
[7] D. Boneh, C. Dunworth, and R. Lipton, *Breaking DES using a molecular computer*, Technical Report CS-TR-489-95, Dept. of Computer Science, Princeton University, 1995.

[8] K. J. Breslauer, R. Frank, H. Blöcker, and L. A. Markey, *Predicting duplex DNA stability from the base sequences*, Proc. National Academy of Sciences USA, 83(11), 3746–3750, June 1986.

[9] A. R. Calderbank and N. J. A. Sloane, *Modular and p-adic cyclic codes*, Designs, Codes, Cryptogr., 6, 1995, 21–35.

[10] H. Dinh and S. R. López-Permouth, *Cyclic and negacyclic codes over finite chain rings*, IEEE Trans. Inform. Theory, 50, 1728–1744, 2004.

[11] A. D’yachkov, A. Macula, T. Renz, P. Vilekin, and I. Ismagilov, *New results on DNA codes*, in Proc. IEEE Int. Symp. Inform. Theory, 283–287, 2005.

[12] A. G. D’yachkov, A. J. Macula, W. K. Pogozelsky, T. E. Renz, V. Rykov and D. C. Torney, A weighted insertion-deletion stacked pair thermodynamic metric for DNA codes, Edit. Ferreti, G. Mauri and C. Zandron DNA10, LNCS 3384, 90–103, 2005.

[13] A. D’yachkov, D. Torney, P. Vilekin and S. White, Reverse-complement similarity codes, Inform. Transfer and Combinatorics, LNCS 4123, Eds. R. Ahlswede et al, 2006.

[14] A. D’yachkov, A. Voronina, A. Macula, T. Renz and V. Rykov, *On critical relative distance of DNA codes for additive stem similarity*, Proc. IEEE Int. Symp. Inform. Theory, 2010.

[15] A. D’yachkov, A. N. Voronina, *DNA codes for additive stem similarity*, Prob. Inform. Transmission 45, 2, 56-77, 2009.

[16] P. Gaborit and H. King, *Linear constructions for DNA codes*, Theoretical Computer Science, 334(1-3), 99–113, 2005.
[17] G. Ganske and B. R. McDonald, *Finite local rings*, Rocky Mountain J. Math. 3(4), 521–540, 1973.

[18] K. Guenda and T. A. Gulliver, Construction of cyclic codes over $\mathbb{F}_2+u\mathbb{F}_2$ for DNA computing, submitted to Appl. Algebra in Eng. Commun. and Computing, Sept. 2011.

[19] K. Guenda and T. A. Gulliver, MDS and self-dual codes over rings, submitted to Finite Fields Appl., 2011.

[20] R. J. Lipton, *DNA solution of hard computational problems*, Science, 268, 542–545, Apr. 1995.

[21] F. J. Macwilliams and N. J. A. Sloane, *The Theory of Error Correcting-Codes*, North-Holland, Amsterdam, 1977.

[22] M. Mansuripur, P. K. Khulbe, P. K khulbe, S. M. Kuebler, J. W. Perry, M. S. Giridhar, and N. Peyghambarian, *Information storage and retrieval using macromolecules as storage media*, University of Arizona Technical Report, 2003.

[23] J. L. Massey, *Reversible codes*, Inform. Control, 7(3), Sep. 1964.

[24] B. R. McDonald, *Finite Rings with Identity*, Pure and Applied Mathematics, 28, New-York, Marcel Dekker, 1974.

[25] O. Milenkovic and N. Kashyap, *On the design of codes for DNA computing*, in Lecture Notes in Computer Science, vol. 3969, Springer-Verlag, 100–119, 2006.

[26] R. Nussinov and A. B. Jacobson, *Fast algorithm for predicting the secondary structure of single stranded RNA*, Proc. Natl. Acad. Sci. USA, 77(11), 6309–6313, 1980.

[27] Q. Ouyang P. D. Kaplan, S. Lin, and A. Libchaber, *DNA solution of the maximal clique problem*, Science, 278, 446–449, 1997.
[28] J. SantaLucia, *A unified view of polymer, dumbell, and oligonucleotide DNA nearest-neighbhor thermodynamics*, Proc. Nat. Acad. Sciences, 95, 1460–1465, 1998.

[29] D. Shoemaker, D. A. Lashkari, D. Morris, M. Mittman, and R. W. Davis, *Quantitative phenotypic analysis of yeast deletion mutant using a highly parallel molecular bar-coding strategy*, Nature Genetics, 16, 450–456, 1996.

[30] I. Siap, T. Abualrub, and A. Ghrayeb, *Cyclic DNA codes over the ring $F_2[u]/(u^2 – 1)$ based on the deletion distance*, J. Franklin Institute, 346, 731–740, 2009.

[31] P. Šulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. K. Doye, and A. A. Louis, *Sequence-dependent thermodynamics of a coarse-grained DNA model*, arXiv:1207.3391v1 [physics.bio-ph].

[32] O. Zariski and P. Samuel, *Commutative Algebra*. New York: Van Nostrand, 1958