On Cyclic DNA Codes

K. Guenda and T. Aaron Gulliver

Dept. of Electrical and Computer Engineering, University of Victoria
PO Box 3055, STN CSC, Victoria, BC Canada V8W 3P6
Email:{kguenda,agullive}@ece.uvic.ca

Patrick Solé

Telecom-ParisTech
CNRS/LTCI
Email: patrick.sol@telecom-paristech.fr

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. Then, cyclic codes over R are designed such that their images under the mapping are also cyclic or quasi-cyclic of index 2 with designed hybridization energy. The hybridization energy and additive distance are shown to be functions of the neighborhood energy.**

I. 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 $3' - ACTTAGA - 5'$ is the strand $5' - TCTAAGT - 3'$. Nucleotide pairing is based on hydrogen bonds, with a pair $A - T$ forming two bonds.

The combinatorial properties of DNA sequences can be used to tackle computationally difficult problems. Adleman et al. [3] used the WCC approach to break the data encryption standard (DES). In addition, Lipton [14] used DNA strands to solve the satisfiability (SAT) problem. Since there are 4^n possible single DNA strands of length n which can quickly and cheaply be synthesized, Mansuripur et al. [15] showed that DNA codewords can be used for ultra high density data storage. Other applications exploit DNA hybridization [20].

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 complement 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.

Secondary structure occurs when a strand folds back onto itself forming a double strand. This phenomena causes codewords to become computationally inactive. Milenkovic and Kashyap [17] used the Nussinov-Jacobson algorithm [18] 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], [10], [12], [22]. The problem of hybridization energy has also been studied extensively [4], [8], [21]. More recently, D'Yachkov et al. [9] modeled the hybridization energy for DNA strands as an additive stem similarity using the neighborhood energy of pairs of nucleotides. They also introduced the additive stem distance. Yildiz and Siap [22] constructed DNA codes as cyclic reversible complement codes of odd length over the ring $R = \mathbb{F}_2[u]/(u^4 - 1)$. They also studied the Hamming distance of these codes.

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 map between the elements of $\mathcal{A}^2 = \{A, C, G, T\}^2$

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-roots 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 one such ring.

Typically the Hamming distance or deletion distance is 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 the DNA codes given here. 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. This is useful in DNA computing since it results in greater stability.

The next section presents some basic facts and preliminary results. 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. Note that another mapping was given in [22], but they did not establish the link between the elements of R and the additive stem similarity or the hybridization energy.

II. PRELIMINARIES

The ring considered here is

$$
\begin{aligned}
R &= \mathbb{F}_2[u]/(u^4 - 1) \\
&= \{a + bu + cu^2 + du^3 \mid a, b, c, d \in F_2, u^4 = 1\},
\end{aligned}
$$

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

$$\langle 0 \rangle = \langle (u+1)^4 \rangle \subsetneq \langle (u+1)^3 \rangle \subsetneq \langle (u+1)^2 \rangle \subsetneq \langle (u+1) \rangle \subsetneq R.$$

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

A map ϕ 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 I. A simple verification gives that for all $x \in R$ we have

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

Note that other mappings can be defined between R and the nucleotide pairs [22]. The mapping ϕ was chosen because it results in a subcode over the alphabet $\{GC, CC, GG, CG\}$.

GG	0	AT	$1+u$	TA	u^2+u^3	CC	$(1+u)^3$
TG	u^2	GT	1	AA	$u+u^2$	TC	$1+u^2+u^3$
AG	u	GC	$1+u^2$	CT	$1+u+u^2$	AC	$1+u+u^3$
GA	u^3	CG	$u+u^3$	TT	$1+u^3$	CA	$u+u^2+u^3$

Since R^n is an R module, a linear code over R of length n is a submodule $\mathcal{C}$ of R^n. Now let $\mathcal{A}$ be any alphabet. Then a code $\mathcal{C}$ over $\mathcal{A}$ is said to be **cyclic** if it is invariant under a cyclic shift, i.e., $(x_{n-1}, x_0, \ldots, x_{n-2}) \in \mathcal{C}$ provided $(x_0, x_1, \ldots, x_{n-2}, x_{n-1})$ is in $\mathcal{C}$. A code is called **quasi-cyclic** of index l if for any $(x_0, x_1, \ldots, x_{n-2}, x_{n-1}) \in \mathcal{C}$ we have $(x_{n-l}, x_{n-l+1}, \ldots, x_0, x_1, x_{n-l-1}) \in \mathcal{C}$. 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 [7], 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 as $x^r = x_{n-1} x_{n-2} \ldots x_1 x_0$, the complement of x as $x^c = \hat{x}_0 \hat{x}_1 \ldots \hat{x}_{n-1}$, and the reverse complement (also called the Watson-Crick complement) as $x^{rc} = \hat{x}_{n-1} \hat{x}_{n-2} \ldots \hat{x}_1 \hat{x}_0$. A code $\mathcal{C}$ is said to be **reverse complement** if for any $x \in \mathcal{C}$, we have $x^{rc} \in \mathcal{C}$. For $\mathcal{A} = \{A, G, C, T\}$, a code $\mathcal{C}$ of length n over the alphabet $\mathcal{A}$ is called reverse complement if the WCC of each codeword $a \in \mathcal{A}^n$ is also in $\mathcal{C}$.

III. ADDITIVE STEM SIMILARITY DISTANCE

In this section, the additive stem similarity introduced by D'yachkov and Voronina [9] 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. 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. $\Delta G(a, b)$ for these pairs at a temperature of $310°$ is given in Table III. For $x = x_1, \ldots, x_n \in \mathcal{A}^n$ and $y = y_1, \ldots, y_n \in \mathcal{A}^n$, define $S_w(x, y) = \sum_{i=1}^{n-1} s_i^w(x, y)$, where

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

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

TABLE II
NEAREST NEIGHBOR THERMODYNAMIC VALUE FOR STACKED PAIRS [4]

Stacked pair $5' \rightarrow 3'/3' \rightarrow 5'$	$\Delta G_{310°C} kcal/mol$
$AA/TT = TT/AA$	1.02
$AC/TG = GT/CA$	1.46
$AG/TG = CT/GA$	1.29
AT/TA	0.88
$CA/GT = TG/AC$	1.46
$CC/GG = GG/GC$	1.83
CG/GC	2.17
$GA/CT = TC/AG$	1.32
GC/CG	2.24
TA/AT	0.60

between x and y, and it satisfies the following properties

$$S_w(x, y) = S_w(y, x) \leq S_w(x, x).$$

The hybridization energy between x and y is [9]

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

Definition 3.1: For $x, y \in \mathcal{A}^n$, the real number

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

is called the additive stem distance between x and y.

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

IV. CYCLIC DNA CODES

Let $\mathcal{C}$ be a linear code over R, $\mathcal{A} = \{A, C, G, T\}$ and $D(., .)$ the additive stem distance given in Definition 3.1. Since the map ϕ defined in Table I 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

$$D(x, y) = D(\phi(x), \phi(y)). \tag{2}$$

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

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

- $\mathcal{C}$ is a cyclic code, i.e., C is an ideal in

$$R_n = R[x]/(x^n - 1);$$

- for any codeword $x \in \mathcal{C}$, $x \neq x^{rc}$ and $x^{rc} \in \mathcal{C}$; and
- $D(x, y) \geq d, \forall x, y \in \mathcal{C}$.

Let $\mathcal{C}$ be an $[n, d]$ cyclic DNA code over R. Then if $s = \max\{S_w(\phi(x), \phi(x)), x \in \mathcal{C}\}$, from (2) and the definition of the additive stem distance we obtain that

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

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

$$E(\phi(x), \phi(y)) \leq s - d, \forall x, y \in \mathcal{C}. \tag{3}$$

A. 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 [7]. We begin by providing some results for codes of odd length.

Lemma 4.2: ([7]) A cyclic code of odd length n over R is an ideal defined as

$$\mathcal{C} = \langle f_0, (u+1)f_1, (u+1)^2 f_2, (u+1)^3 f_3 \rangle, \tag{4}$$

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

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

$$(-) : R[x] \longrightarrow \mathbb{F}_2[x]$$
$$f \longmapsto \overline{f} = f \mod u + 1. \tag{5}$$

Definition 4.3: A polynomial f in $R[x]$ is called regular if $\overline{f} \neq 0$. 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 4.4: ([16, Theorem XIII. 11]) Let f be a regular polynomial in $R[x]$. Then $f = \alpha g_1 \ldots g_r$ where α 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 α, β 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 4.5: 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 $k > 3$, we have $g^k \in \langle f(x) \rangle$, and thus $f(x)$ is a primary polynomial. □

Remark 4.6: Let m be an odd integer. Then from [13], the polynomial $x^m - 1$ factors uniquely as a product of

3

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

$$x^n - 1 = g_1^{2^s} \ldots g_l^{2^s}, \tag{6}$$

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 $2|\binom{2^s}{i}$ for $1 \leq i \leq 2^s$. From Remark 4.6, $x^m - 1$ 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 $x^n - 1 = f_1^{\alpha_1} \ldots f_r^{\alpha_l}$ is a decomposition into powers of basic irreducible polynomials. From Lemma 4.5 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 4.4, the decomposition (6) is unique. $\square$

Proposition 4.8: With the previous notation, the primary ideals of $R[x]/(x^n - 1)$ are $\langle 0 \rangle$, $\langle 1 \rangle$, $\langle g_i^j \rangle$, $\langle g_i^j, (u+1)^t \rangle$, with $1 \leq j \leq 2^s$, $1 \leq t \leq 3$ and $1 \leq i \leq l$.

Proof. Let $\mu : R[x] \longmapsto \frac{\mathbb{F}_2[x]}{\langle x^n - 1 \rangle}$ be the canonical homomorphism. By Lemma 4.7, 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 μ is the ideal $\langle x^n - 1, u \rangle$. Hence from [23, Theorem 3.9.14], the primary ideals of $R[x]/(x^n - 1)$ are the preimages of the primary ideals of $\mathbb{F}_2[x]/(x^n - 1)$. It is well known [11, Theorem 3.10] that the primary ideals of this last ring are the ideals $\langle g_i^j \rangle$, $1 \leq j \leq 2^s$, $1 \leq i \leq l$. Hence the primary ideals of $\mathcal{R}$ are $\langle g_i^j, (u+1)^t \rangle$. $\square$

Theorem 4.9: Let $n = m2^s$ such that m is an odd integer. Then the cyclic codes of length $m2^s$ over R are the ideals generated by $\langle f_0, (u+1)f_1, (u+1)^2 f_2, (u+1)^3 f_3 \rangle$, where $f_i | f_0 | x^n - 1$, for $1 \leq i \leq 3$.

Proof. Let $\mathcal{C}$ be a cyclic code in $R[x]$ so that $\mathcal{C}$ is an ideal of $\mathcal{R}$. Since $\mathcal{R}$ is Noetherian, from the Lasker-Noether Decomposition Theorem [23, p. 209], any ideal in $\mathcal{R}$ has a representation as a product of primary ideals. From Proposition 4.8, we have that the primary ideals of $\mathcal{R}$ are $\langle g_i^j, (u+1)^t \rangle$, where $x^n - 1 = \prod_{l=1}^r g_i^{2^s}$. Hence an ideal I of $\mathcal{R}$ is of the form

$$I = \prod_{l=1}^r \langle g_i^j, (u+1)^t \rangle. \tag{7}$$

Expanding the product in (7), each ideal in $\mathcal{R}$ is generated by

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

where $f_i | f_0 | x^n - 1$, for $1 \leq i \leq 3$. $\square$

B. 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 4.10 ([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) $[f(x)g(x)]^* = f(x)^* g(x)^*$; and

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

Theorem 4.11: Let $\mathcal{C}$ be a reverse-complement cyclic code over R. Then the following holds:

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

(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 [22]. Part (ii) in the case n odd was proven in [22, Theorem 4.3]. Since by Theorem 4.9 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 [22] for n odd also holds for the case n even. $\square$

The proof of the following theorem is the same as that of [22, Theorem 4.4] for odd length.

Theorem 4.12: Let $\mathcal{C}$ be a cyclic code over R of length n. Suppose $(1 + u + u^2 + u^3)\mathbb{I}(x) \in \mathcal{C}$. Then if $\mathcal{C} = \langle f_0, (u+1)f_1, (u+1)f_1^2, (u+1)^3 f_3 \rangle$, where the f_i are self-reciprocal, then $\mathcal{C}$ is a reverse-complement code.

Corollary 4.13: Let $\mathcal{C}$ be a cyclic code of length $n = m2^s$, $s \geq 0$ and m odd. Then if $(1 + u + u^2 + u^3)\mathbb{I}(x) \in \mathcal{C}$ and if there exists an i such that

$$2^i \equiv -1 \mod m, \tag{8}$$

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 (8) gives that $Cl(1)$ is reversible and hence all the cyclotomic classes are reversible. Thus each minimal polynomial is self-reciprocal, and from Lemma 4.10 the polynomials f_i are self-reciprocal. Then from Theorem 4.12, $\mathcal{C}$ is a reverse-complement code. $\square$

Example 4.14: 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 reverse-complement code over R.

Corollary 4.15: 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 $\{A, G, C, T\}$.

Proof. Let $\mathcal{C}$ be a cyclic DNA code of length n over R. Hence $\phi(\mathcal{C})$ is a set of length $2n$ over the alphabet $\mathcal{A}$ which is quasi-cyclic of index 2. Since $\mathcal{C}$ is a reverse-complement code, then $u^2 x^{rc} \in \mathcal{C}$, and $\phi(u^2 x^{rc})$ is the WCC of $\phi(x)$. $\square$

Definition 4.16: For a code $\mathcal{C}$ over R, define the subcode $\mathcal{C}_{1+u^2}$ consisting of all codewords in $\mathcal{C}$ that are a multiple of $(1 + u^2)$.

Lemma 4.17: With the previous definition we have

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

Further, if $\mathcal{C} = \langle f_0, (u+1)f_1, (u+1)^2 f_2, (u+1)^3 f_3 \rangle$ is a cyclic code of length n over R, if we assume that $f_3 | f_2 | f_1 | f_0$, then

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

Proof. The first part of the lemma is a simple verification. For the second part, assume $\mathcal{C} = \langle f_0, (u+1)f_1, (u+1)^2 f_2, (u+1)^3 f_3 \rangle$. Since $f_3 | f_2 | f_1 | f_0$, we then have $\langle (1 + u^2) f_3 \rangle \subset \mathcal{C}_{1+u^2}$. Now let $c(x) \in \mathcal{C}$ so that $c(x) = k_0(x) f_0(x) + (u+1) k_1(x) f_1(x) + (u+1)^2 k_2(x) f_2(x) + (u+1)^3 k_3(x) f_3(x)$ for $k_i(x) \in \mathbb{F}_2[x]$. If $c(x)$ is a multiple of $1 + u^2$, then $x^n - 1 | k_0(x) f_0(x)$ 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 $f_3(x) | f_2(x) | f_1(x)$, $\mathcal{C}_{1+u^2} \subset \langle (1 + u^2) f_3(x) \rangle$, therefore $\mathcal{C}_{1+u^2} = \langle (1 + u^2) f_3 \rangle$. $\square$

Theorem 4.18: Let $\mathcal{C} = \langle f_0, (u + 1)f_1, (u + 1)^2 f_2 | (u + 1)^3 f_3 \rangle$ be an $[n, d]$ cyclic DNA code over R such that $f_3 | f_2 | f_1 | f_0$. Let $d_{1+u^2} = \min\{D(x, y), \ x, y \in \mathcal{C}_{1+u^2}\}$. Then $\phi(\mathcal{C}_{1+u^2})$ 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 $\mathcal{C}_{1+u^2} \subset \mathcal{C}$, it is obvious that $d_{1+u^2} \geq d$. From Theorem 4.15, we have that the image of the cyclic DNA code $\mathcal{C}$ obtained via ϕ is a quasi-cyclic code of length $2n$ over the alphabet $\{A, G, C, T\}$. From Lemma 4.17, $\phi((1 + u^2)R) = \{GG, CC, CG, GC\}$ and $\mathcal{C}_{1+u^2} = \langle (1 + u^2) f_3(x) \rangle$, which gives the result. $\square$

Remark 4.19: Theorem 4.18 is useful as it results in cyclic subcodes with large GC-content. Since from Table I the stacked pair corresponding to $\{GG, CC, GC, CG\}$ has the largest neighborhood energy, these subcodes will have high hybridization energy.

REFERENCES

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

[4] 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.

[5] D. Boneh, C. Dunworth, and R. Lipton, *Breaking DES using a molecular computer,* Technical Report CS-TR-489-95, Dept. of Computer Science, Princeton Univesity, 1995.

[6] 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, Jun. 1986.

[7] H. Dinh and S. R. López-Permouth, *Cyclic and negacyclic codes over finite chain rings*, IEEE Trans. Inform. Theory, 50(8), 1728–1744, Aug. 2004.

[8] 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, 1325–1329, Jun. 2010.

[9] A. D'yachkov and A. N. Voronina, *DNA codes for additive stem similarity*, Prob. Inform. Transmission, 45(2), 124–144, 2009.

[10] P. Gaborit and H. King, *Linear constructions for DNA codes*, Theoretical Computer Science, 334(1-3), 99–113, 2005.

[11] G. Ganske and B. R. McDonald, *Finite local rings*, Rocky Mountain J. Math. 3(4), 521–540, 1973.

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

[13] K. Guenda and T. A. Gulliver, *MDS and self-dual codes over rings*, Finite Fields and their Applic., 18(6), 1061-1075, Nov. 2012.

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

[15] M. Mansuripur, 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.

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

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

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

[19] J. SantaLucia, *A unified view of polymer, dumbell, and oligonucleotide DNA nearest-neighbhor thermodynamics*, Proc. Nat. Acad. Sciences, 95(4), 1460–1465, Feb. 1998.

[20] 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.

[21] 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(8), 731–740, Oct. 2009.

[22] B. Yildiz and I. Siap, *Cyclic DNA codes over $F_2[u]/(u^4 - 1)$ and applications to DNA codes*, Com. Math App. 63(7), 1169–1176, Apr. 2012.

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