ROMAI J., v.10, no.2(2014), 203–211

A METHOD FOR CONSTRUCTING DNA CODES FROM ADDITIVE SELF-DUAL CODES OVER *GF*(4)

Zlatko Varbanov[1], Todor Todorov[1,2], Maya Hristova[1]
[1] *Dept. of Information Technologies, University of Veliko Tarnovo, Bulgaria*
[2] *Institute of Mathematics and Informatics, Bulgarian Academy of Sciences*
zl.varbanov@uni-vt.bg, todorvt@abv.bg, maqhristova@yahoo.com

Abstract In this paper we translate in terms of coding theory constraints that are used in designing DNA codes for use in DNA computing. We focus in particular on additive self-dual codes over $GF(4)$, and we propose a new method for constructing DNA codes satisfying the Hamming distance constraint, the reverse-complement constraint and the GC-content constraint.
Acknowledge. This research is supported by RD672-08/2012 Project, University of Veliko Tarnovo.
Work presented at CAIM 2014, September 19-22, "Vasile Alecsandri" University of Bacău, Romania.

Keywords: DNA codes, additive circulant graph codes, constraints on DNA codes.
2010 MSC: 94B60, 94B65.

1. INTRODUCTION

Coding theory has several applications in Genetics and Bioengineering. Every DNA molecule consists of two complementary strands which are sequences of four different nucleotide bases called adenine (A), cytosine (C), guanine (G) and thymine (T). The problem of designing DNA codes (sets of words of fixed length n over the alphabet $\{A, C, G, T\}$ that satisfy certain combinatorial constraints has applications for reliably storing and retrieving information in synthetic DNA strands.

In this work we translate in terms of coding theory [4, 5] the constraints that are used in designing DNA codes for use in DNA computing. In [1, 3, 6] four different constraints are considered: the Hamming distance constraint, the reverse-complement constraint, the reverse constraint and the fixed GC-content constraint. We propose new construction for DNA codes satisfying a Hamming distance constraint, the reverse-complement constraint and the fixed GC-content constraint. Practically, the focus is on additive self-dual codes over $GF(4)$ and their graph representation.

The paper is organized as follows: Section 2 recalls basic notions for DNA codes and additive codes over $GF(4)$. In Section 3 the constraints on DNA codes are translated into the terms of additive codes. Section 4 lists some known constructions

and presents our new construction. In the end of Section 4 we give a table with the obtained results for DNA codes satisfying the Hamming distance constraint, the reverse-complement constraint and the GC-content constraint.

2. DNA CODES AND ADDITIVE CODES OVER $GF(4)$

Definition 2.1. *[3] A DNA code of length n is a set of codewords* $(x_1, \ldots, x_n)$ *with* $x_i \in \{A, C, G, T\}$ *(representing the four nucleotides in DNA). We use a hat to denote the Watson-Crick complement of a nucleotide, so* $\hat{A} = T, \hat{T} = A, \hat{C} = G$, *and* $\hat{G} = C$.

Definition 2.2. *The Hamming distance* $H(x, y)$ *between two codewords x and y is the number of coordinates in which x and y differ.*

Definition 2.3. *The reverse of a codeword* $x = (x_1, \ldots, x_n)$ *is denoted by* $x^R = (x_n, \ldots, x_1)$, *and the reverse-complement of x is denoted by* $x^{RC} = (\hat{x}_n, \ldots, \hat{x}_1)$.

In this paper we shall identify codes over $\{A, C, G, T\}$ with codes over other four-letter alphabet. In our case this is the field $GF(4) = \{0, 1, \omega, \omega^2\}$, with $\omega^2 + \omega + 1 = 0$. The four symbols in $\{A, C, G, T\}$ are identified with the four symbols in $GF(4)$ in the following way: $0 \to A$, $1 \to T$, $\omega \to C$, and $\omega^2 \to G$, so that $\hat{x} = x + 1$, for $x \in GF(4)$.

Let $GF(4)^n$ be the n-dimensional vector space over the Galois field $GF(4)$. The *Hamming weight* of a vector $x \in GF(4)^n$, written $wt(x)$, is the number of nonzero entries of x.

Definition 2.4. *A quaternary additive* $(n, 2^k)$ *code of length n is an additive subgroup of* $GF(4)^n$ *with* 2^k *codewords (where* $0 < k \leq n$*).*

A *minimum weight* (or *minimum distance*) of an additive code is the smallest weight among all nonzero codewords. By $(n, 2^k, d)$ we denote an additive code of length n with 2^k codewords that has minimum distance d.

Definition 2.5. *Any* $k \times n$ *matrix G (*$0 < k \leq n$*) with entries in* $GF(4)$ *whose rows are a basis of the code C is a generator matrix of C.*

About additive codes over $GF(4)$, there is an inner product arising from the trace map $Tr : GF(4) \to GF(2)$ given by $Tr(x) = x + x^2$. The conjugate of $x \in GF(4)$, denoted $\bar{x}$, is the following image: $\bar{0} = 0, \bar{1} = 1$, and $\bar{\omega} = \omega^2$. Now we can define the trace inner product of two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $GF(4)^n$ as:

$$x * y = \sum_{i=1}^{n} Tr(x_i . \bar{y}_i) \tag{1}$$

Definition 2.6. *The dual code of an additive code C (with respect to (1)) is the code* $C^{\perp} = \{x \in GF(4)^n | x * c = 0$ *for all* $c \in C\}$. *If C is an* $(n, 2^k)$ *code, then* $C^{\perp}$ *is an* $(n, 2^{2n-k})$ *code. The code C is self-orthogonal if C is a subset of* $C^{\perp}$, *and self-dual if* $C = C^{\perp}$.

3. CONSTRAINTS ON DNA CODES

- *Hamming distance constraint*: the Hamming distance constraint for a DNA code C is that $H(x, y) \geq d$ for all $x, y \in C$ with $x \neq y$, for some prescribed minimum distance d. This constraint will be enforced in all of the codes we consider, in addition to some combination of the constraints described below.
- *Reverse constraint*: the reverse constraint is that $H(x^R, y) \geq d$ for all $x, y \in C$, including $x = y$. It is useful as an intermediate step in constructing codes with the reverse-complement constraint.
- *Reverse-complement constraint* (RC-constraint): this constraint is that $H(x^{RC}, y) \geq d$ for all $x, y \in C$, including $x = y$. To construct codes satisfying the reverse-complement constraint, it can be useful to begin with codes over $GF(4)$ that contain a special codeword we denote by $\tilde{1}$, which is the all-one word for $GF(4)$. Note that $x^{RC} = x^R + \tilde{1}$, so an additive code containing $\tilde{1}$ that is fixed by the permutation $x \to x^R$ is also fixed by the map $x \to x^{RC}$.
- *GC-content constraint*: this constraint is that each codeword $x \in C$ has the same GC-weight (the number of entries of the codeword that are C or G).

Starting from an additive code, the question is how to compute the GC-weight of all codewords. This turns out to be quickly time consuming, since finding of all codewords may in itself take a long time. In this work we propose a simple way to compute the number of codewords with fixed GC-weight of a special class of additive self-dual codes over $GF(4)$. Also, it is easy to determine the number of codewords (among the codewords with fixed GC-weight) that satisfy the RC-constraint.

4. CONSTRUCTIONS ON DNA CODES

Tables with lower bounds on DNA codes can be found in [3] and [7]. Also, in [3] and [7] some constructions on DNA codes can be found:

- Binary construction
- Lexicographic construction
- Linear reverse construction
- Cyclic (and extended cyclic) code construction
- Shortening and puncturing

In our work we present a construction method based on additive self-dual codes with circulant generator matrix in graph form. The corresponding definitions are given below. This method was already used to construct DNA codes that satisfy the Hamming distance constraint and the GC-weight constraint [9]. Here we use the

method to construct DNA codes that satisfy these two constraints and also satisfy the reverse-complement constraint.

Definition 4.1. *A graph code is an additive self-dual code over $GF(4)$ with generator matrix $G = \Gamma + \omega I$ where I is the identity matrix and Γ is the adjacency matrix of a simple undirected graph.*

Example:

$$\Gamma = \begin{pmatrix} 0\ 1\ 1 \\ 1\ 0\ 1 \\ 1\ 1\ 0 \end{pmatrix}, \quad C = \Gamma + \omega I = \begin{pmatrix} \omega\ 1\ 1 \\ 1\ \omega\ 1 \\ 1\ 1\ \omega \end{pmatrix}$$

Schlingemann [8] first proved (in terms of quantum stabilizer states) that for any self-dual quantum code, there is an equivalent graph code. This means that there is a one-to-one correspondence between the set of simple undirected graphs and the set of additive self-dual codes over $GF(4)$.

A matrix B of the form

$$B = \begin{pmatrix} b_0 & b_1 & \dots & b_{n-2} & b_{n-1} \\ b_{n-1} & b_0 & b_1 & \dots & b_{n-2} \\ \dots & \dots & \dots & \dots & \dots \\ b_2 & \dots & b_{n-1} & b_0 & b_1 \\ b_1 & b_2 & \dots & b_{n-1} & b_0 \end{pmatrix}$$

is called a circulant matrix. The vector $(b_0, b_1, \dots, b_{n-1})$ is called a *generator vector* for the matrix B. An additive self-dual code with circulant generator matrix is called a *circulant code*.

Definition 4.2. *An additive circulant graph (ACG) code is a code corresponding to a graph with circulant adjacency matrix.*

Circulant graphs must be regular, i.e., all vertices must have the same number of neighbors.

Proposition 4.1. *The generator vector g of a circulant adjacency matrix G has the following property: $g_i = g_{n-i}$, for all $i = 1, \dots, n-1$, and $g_0 = \omega$.*

Proof. Let G be a circulant adjacency matrix and let $g = (g_0, g_1, \dots, g_{n-1})$ be its generator vector. It is known that $g_0 = \omega$ because G is a generator matrix of a graph code. The second entry in the first column of G is equal to g_{n-1} (because the matrix is circulant) but also it is equal to g_1 (because the matrix is symmetric), then $g_{n-1} = g_1$. Analogously, we obtain $g_{n-2} = g_2$, $g_{n-3} = g_3$, etc. Therefore $g_i = g_{n-i}$, for all $i = 1, \dots, n-1$. ∎

Then, the entries of this matrix depend on the coordinates $(g_1, g_2, \dots, g_{\lfloor n/2 \rfloor})$ only. Therefore, we can restrict our search space to the $2^{\lfloor n/2 \rfloor}$ codes over $GF(4)$ of length n corresponding to graphs with circulant adjacency matrices.

Besides a smaller search space, the special form of the generator matrix of a graph code makes it easier to determine the minimum distance, since any codeword obtained as a binary linear combination of i different rows of the generator matrix must have weight at least i (because in any row and column there is just one position that is neither 0 nor 1). Then, if we want to check whether a code C has minimum distance at least d, we only need to consider the combinations of $d-1$ or fewer rows of its generator matrix.

Let $A_4^{GC}(n,d,u)$ denote the maximum size (the maximum number of codewords) of a DNA code of length n with fixed GC-content u that satisfies the Hamming distance constraint for a given d and let $A_4^{GC,RC}(n,d,u)$ denote the maximum size of a DNA code of length n with fixed GC-content u that satisfies the Hamming distance constraint and the RC-constraint for a given d. In our previous work [9] we have proved that graph codes are proper to construct DNA codes with fixed GC-content u that satisfy Hamming distance constraint for given d. The following theorem was proved in [9].

Theorem 4.1. *Any ACG code of length n with minimum distance d consists of DNA codes of length n with* $H(x,y) \geq d$*, fixed GC-content u* $(1 \leq u \leq n)$*, and* $A_4^{GC}(n,d,u) = \binom{n}{u}$.

In this work we consider another special properties (useful about the RC-constraint) of the generator matrix G of an ACG code. The i^{th} row is a reverse of the $(n-i+1)^{th}$ row, for any $1 \leq i \leq n/2$. Let R be a reverse permutation (a permutation that exchanges column i and column $n-i+1$ of the code, for $1 \leq i \leq n$). Then any codeword that is a linear combination of even number of pairwise reversed rows of G is fixed by R. Also, any ACG code contains all-ω or all-$\bar{\omega}$ codeword (that is the linear combination of all rows of G). Therefore, it is equivalent to a code that contains the codeword $\tilde{1}$ (by multiplication of all columns of G by $\bar{\omega}$ or ω, respectively). More information about the equivalence of additive codes over $GF(4)$ can be found in [2].

Example 1: Let $n = 7$ and the generator vector of a circulant adjacency matrix be $(\omega 101101)$. Then the 2^{nd} row is $(1\omega 10110)$ and the 6^{th} row is $(01101\omega 1)$. The sum of these two reversed rows is the codeword $(1\bar{\omega}000\bar{\omega}1)$ that is fixed by R. Any sum of codewords fixed by R is also a codeword fixed by R.

Let C be a code fixed by R. Then C can be written as a disjoint union $C = C_0 \cup C_1 \cup C_2$, where C_0 is the set of codewords in C that are unchanged by R, and C_1 and C_2 are two sets that are interchanged by R [3]. A set satisfying the reverse constraint together with the Hamming constraint d is obtained by taking C_1 or C_2. Obviously, $|C_1| = \frac{|C|-|C_0|}{2}$.

Any ACG code is fixed by a reverse permutation R (because of the symmetric form of its generator matrix).

Proposition 4.2. *Let C be an ACG code of even length n with generator matrix G and let R be a reverse permutation. Then a codeword* $x \in C$ *belongs to the set* C_0 *if and*

only if this codeword is a linear combination of even number of pairwise reversed rows of G.

Proof. : If the length n is even the matrix G does not contain a row that is fixed by R (because any row of G contains only one element ω). Then if the codeword x is a sum of two reversed rows (see Example 1) it is fixed by R. Any linear combination of codewords fixed by R is also a codeword fixed by R. In this way we obtain that the codewords unchanged by R (the set C_0) are these codewords that are linear combinations of pairwise reversed rows of G. ∎

Therefore, for odd u we can take the set of codewords of the corresponding DNA code with fixed GC-weight u and this set does not contain a codeword belongs to C_0 (because all of these codewords are linear combinations of odd number of rows of G).

Using the result in Theorem 4.1 we obtain the following:

Theorem 4.2. *Any ACG code of even length $n \equiv 2$ (mod 4) with minimum distance d consists of DNA code of length n with $H(x, y) \geq d$, fixed GC-content $u = n/2$, and $A_4^{GC,RC}(n, d, u) = \binom{n}{u}/2$.*

Proof. Any ACG code is equivalent to a code that contains $\tilde{1}$. If $n \equiv 2$ (*mod* 4) and $u = n/2$ then n is even and u is odd. The corresponding DNA code does not contain codewords fixed by a reverse permutation R. Then $|C_0| = 0$ and $|C_1| = |C_2| = |C|/2$. By Theorem 4.1 there are $\binom{n}{u}$ codewords with fixed GC-weight u, therefore $A_4^{GC,RC}(n, d, u) = |C_1| = \binom{n}{u}/2$. ∎

The ACG codes used in all examples below were constructed in [10].

Example 2: Let $n = 22$ and $u = 11$. There exists $(22, 2^{22}, 8)$ ACG code. Then $A_4^{GC,RC}(22, 8, 11) = \binom{22}{11}/2 = 352716$ (in [7] this value is 353496).

Example 3 (new lower bound): Let $n = 30$ and $u = 15$. There exists $(30, 2^{30}, 12)$ ACG code. Then $A_4^{GC,RC}(30, 12, 15) = \binom{30}{15}/2 = 77558760$ (in [7] this value is 281928).

Let C be an ACG code of length $n \equiv 0$ (*mod* 4) (i.e. $n = 4k$ $(k > 0)$) with generator matrix G. Then the value $u = n/2$ is even and there are codewords fixed by a reverse permutation R because for any row among the first $2k$ rows of G we can take the corresponding reversed row among the second $2k$ rows of G. Their combination is a codeword fixed by R. We have $u = 2k$ therefore if we take any k rows among the first $2k$ rows and the corresponding k reversed rows from the second $2k$ rows, then their combination is a codeword fixed by R. There are $\binom{2k}{k}$ such codewords and then $|C_0| = \binom{2k}{k}$. All other combinations are codewords that are not fixed by R. In this way we prove the following:

Theorem 4.3. *Any ACG code of even length* $n = 4k$ $(k > 0)$ *with minimum distance* d *consists of DNA code of length* n *with* $H(x, y) \geq d$*, fixed GC-content* $u = n/2 = 2k$*, and* $A_4^{GC,RC}(n, d, u) = \frac{\binom{n}{2k} - \binom{2k}{k}}{2}$.

Example 4: Let $n = 20$ and $u = 10$. There exists $(20, 2^{20}, 8)$ ACG code. Then $A_4^{GC,RC}(20, 8, 10) = (\binom{20}{10} - \binom{10}{5})/2 = 92252$ (in [7] this value is 184756).

If G is the generator matrix of an ACG code C of odd length n, this matrix contains one row fixed by a reverse permutation R.

Proposition 4.3. *Let* C *be an ACG code of odd length* n *with generator matrix* G *and let* R *be a reverse permutation. If* u *is the fixed GC-weight, then a codeword belongs to the set* C_0 *if this codeword is a linear combination of even number of pairwise reversed rows of* G *(for even* u*) or a linear combination of even number of reversed rows of* G *and the row fixed by* R *(for odd* u*).*

Proof. Analogously to Proposition 4.2. ∎

Therefore, if $n = 4k + 1$ in both cases (for $u = 2k$ and for $u = 2k + 1$) the set C_0 contains the same number of codewords. If we take even GC-weight $u = 2k$ then $|C_0| = \binom{2k}{k}$.

Theorem 4.4. *Any ACG code of odd length* $n = 4k + 1$ *with minimum distance* d *consists of DNA code of length* n *with* $H(x, y) \geq d$*, fixed GC-content* $u = 2k$*, and* $A_4^{GC,RC}(n, d, u) = \frac{\binom{n}{2k} - \binom{2k}{k}}{2}$.

Example 5: Let $n = 17$ and $u = 8$. There exists $(17, 2^{17}, 8)$ ACG code. Then $A_4^{GC,RC}(17, 8, 8) = (\binom{17}{8} - \binom{8}{4})/2 = 12120$ (the same as the result in [7]).

For $n = 4k + 3$ we take $u = 2k + 1$. In this case $|C_0| = \binom{2k+1}{k}$ and obtain the following:

Theorem 4.5. *Any ACG code of odd length* $n = 4k + 3$ *with minimum distance* d *consists of DNA code of length* n *with* $H(x, y) \geq d$*, fixed GC-content* $u = 2k + 1$*, and* $A_4^{GC,RC}(n, d, u) = \frac{\binom{n}{2k+1} - \binom{2k+1}{k}}{2}$.

Example 6: Let $n = 23$ and $u = 11$. There exists $(23, 2^{23}, 8)$ ACG code. Then $A_4^{GC,RC}(23, 8, 11) = (\binom{23}{11} - \binom{11}{5})/2 = 675808$ (in [7] this result is 676312).

All obtained results are summarized in the following theorem.

Theorem 4.6. *Any ACG code of length* n *with minimum distance* d *consists of DNA code of length* n *with* $H(x, y) \geq d$*, fixed GC-content* $u = \lfloor n/2 \rfloor$ *and* $A_4^{GC,RC}(n, d, u) = (\binom{n}{u} - \binom{u}{k})/2$*, except in the case* $n \equiv 2 \pmod 4$*, where* $A_4^{GC,RC}(n, d, u) = \binom{n}{u}/2$.

n	d	our result	bound in [7]	n	d	our result	bound in [7]
13	5	848	1696	25	8	2599688	10399676
14	6	1716	3712	26	8	5200300	83204800
15	6	3200	6648	27	8	10028292	40114884
16	6	6600	55424	28	10	20056584	80226336
17	7	*12120*	12120	29	11	*38777664*	38777664
18	6	24310	699624	30	12	**77558760**	281928
19	7	46126	92252	31	10	**150266880**	–
20	8	92252	184756	32	10	**300533760**	–
21	7	176232	176772	33	10	**583395120**	–
22	8	352716	353496	34	10	**1166803110**	–
23	8	675808	676312	35	10	**2268771670**	–
24	8	1351616	5406464	36	10	**4537543340**	–

Table 1: Lower bounds on $A_4^{GC,RC}(n,d,u)$, $13 \leq n \leq 36$ obtained by additive circulant graph codes

Table 1 consists of the results for DNA codes obtained by the ACG codes with maximum reached d constructed in [10]. Entries in bold face denote new best lower bounds. Entries in italics are codes which match the best known values as given in [7].

5. CONCLUSION

In this work we have presented some connections between DNA codes and additive codes over a field with 4 elements. We use a special class of additive self-dual codes and we propose a new construction method for DNA codes based on the form of generator matrices of the codes in this special class. By this construction we obtain some results on DNA codes that satisfy the Hamming distance constraint, the GC-content constraint, and the reverse-complement constraint for given parameters. In some cases our results match the best known values for the maximum size of a DNA code of length n that satisfy these constraints and for $n \geq 30$ we obtain new best lower bounds.

Acknowledgment We would like to thank to the anonymous referees for their helpful remarks.

References

[1] A. G. Frutos, Q. Liu, A. J. Thiel, A. M. W. Sanner, A. E. Condon, L.M. Smith and R. M. Corn, *Demonstration of a word design strategy for DNA computing on surfaces*, Nucleic Acids Research, vol. **25** (1997), 4748–4757.

[2] P. Gaborit, W. C. Huffman, J. L. Kim, V. Pless, *On additive GF(4)-codes*, DIMACS Workshop Codes Assoc. Schemes, DIMACS Ser. Discr. Math. Theoret. Comp. Sci., Amer. Math. Soc. **56** (2001), 135–149.

[3] P. Gaborit, O. D. King, *Linear constructions for DNA codes*, Theoretical Computer Science. **334** (2005), 99–113.

[4] W. C. Huffman and V. Pless, *Fundamentals of coding theory*, Cambridge University Press, 2003.

[5] F. J. MacWilliams and N. J. A. Sloane, *The Theory of error correcting codes*, North Holland, 1977.

[6] A. Marathe, A. E. Condon and R. M. Corn, *On Combinatorial DNA word design*, Journal of Computational Biology, vol. **8** (2001) 201–220.

[7] A. A. Niema, *The construction of DNA codes using a computer algebra system*, PhD Thesis, University of Glamorgan, 2011

[8] D. Schlingemann, *Stabilizer codes can be realized as graph codes*, Quantum Inf. Comput. **2** (2002), 307–323, arXiv:quant-ph/0111080.

[9] T.Todorov, Z.Varbanov, *DNA codes based on additive self-dual codes over GF(4)*, Proc. 7th Int. Workshop on Optimal Codes and Related Topics, Albena, Bulgaria (2013), 170–175

[10] Z. Varbanov, *Additive self-dual graph codes over GF(4)*, Proc. 6th Int. Workshop on Optimal Codes and Related Topics, Varna, Bulgaria (2009), 189–195