Construction of codes for DNA computing by the Greedy Algorithm

Nabil Bennenni, Kenza Guenda and T. Aaron Gulliver
University of Science and Technology, USTHB, Algiers, Algeria.

ken.guenda@gmail.com

Abstract

In this paper we construct codes for DNA computing using the greedy algorithm over $\mathbb{Z}_4$. We obtain linear codes over $\mathbb{Z}_4$ with bounded GC content. We also consider the edit distance, we gave upper bounds for the edit distance and construct codes with bounded edit distance.

Keywords
DNA codes, GC-content, edit distance, upper bound

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), 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, known as base pairing, occurs when a strand binds to another strand, forming a double strand of DNA.

The strands are linked following the Watson-Crick model. 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 $\hat{C} = G$. The 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 $5' - TCTAAGT - 3'$. The WCC property of the DNA strands is used in DNA computing. Namely the data is encoded using DNA strands, and molecular biology techniques are used to simulate arithmetic and logical operations. The main advantages of this approach are huge memory capacity, massive parallelism, and low power molecular hardware and software systems. Other application make use of the DNA properties [7]. In this paper we construct codes for DNA computing using the greedy algorithm over $\mathbb{Z}_4$ [3]. We obtain linear codes over $\mathbb{Z}_4$ with bounded GC content. We also consider the edit distance, we gave upper bounds for the edit distance and construct codes with bounded edit distance. Our choice of the ring $\mathbb{Z}_4$ comme from the fact that the property bounded GC-content and bounded edit distance are multiplicative property over $\mathbb{Z}_4$. This is not the case over $\mathbb{F}_4$. The bounded GC-constraint ensures that all codewords have their thermodynamic characteristics below some threshold. This is an important criteria of these DNA sequences; since it means that they do not produces erroneous crosshybridization. Also the edit distance is an important combinatorial property of the DNA strand. Note the algorithm of Chee and Ling [2] given in order to construct DNA codes with large GC-content is optimal only up to $n = 12$. Bishop et al. [1] considered the construction of fixed GC-content codes using a probabilistic model and random codes. This paper is organized as follows. In Section 2 we give some preliminaries. In Section 3 we give the greedy algorithm for bounded GC-content. In Section 4 we construct DNA lexicodes with edit distance criteria and we give upper bound on the edit distance. Several examples of DNA codes with bounded GC-content and edit distance criteria.

2 Preliminaries

The ring considered here is the ring $\mathbb{Z}_4$ with element $\{0, 1, 2, 3\}$ and the addition and multiplication modulo 4. It is a finite chain ring with maximal ideal $< 2 >$ and nilpotency index 2. We define the Hamming weight of a codeword x in $\mathbb{Z}_4^n$ as $w_H(x) = n_1(x)+n_2(x)+n_3(x)$, the Hamming distances $d_H(\mathsf{x}, \mathsf{y})$, between two vectors x and y is $wt_H(\mathsf{x} - \mathsf{y})$.

2.1 Construction of Lexicode over $\mathbb{Z}_4$

In this section we recall the construction of lexicodes over $\mathbb{Z}_4$ given in [3]. A linear code $\mathcal{C}$ of the length n over $\mathbb{Z}_4$ is an additive over $\mathbb{Z}_4^n$. $\mathbb{Z}_4^n$ is linear code over $\mathbb{Z}_4$ with basis $B = \{b_1 \cdots b_n\}$. With respect to this basis we recursively define a lexicographically ordered list $V_i = x_1, x_2, \cdots, x_{4^i}$ as follows

$$V_0 := 0$$

$$V_i := V_{i-1}, b_i + V_{i-1}, 2b_i + V_{i-1}, 3b_i + V_{i-1}, 1 \leq i \leq n.$$

In this way $|V_i| = 4^i$, and we can identify ${\mathbb{Z}_4}^n$ by V_n. Assume now that we have a property P which can test if a vector $c \in R^n$ is selected or not. That selection property P on V can be seen as a boolean valued function $P : V \rightarrow \{True, False\}$ that depends on one variable. Over $\mathbb{Z}_4$, the property P is called a multiplicative property if $P[x]$ is true implies $P[3x]$ is true.

The following greedy algorithm provides lexicodes over $\mathbb{Z}_4^n$, see article [3].

2.1.1 Algorithm

1. $\mathcal{C}_0 := 0; i := 1;$
2. select the first vector $a_i \in V_i \backslash V_{i-1}$ such that $P[2a_i + c]$ for all $c \in \mathcal{C}_{i-1}$;
3. if such an a_i exists, then $\mathcal{C}_i := \mathcal{C}_{i-1}, a_i + \mathcal{C}_{i-1}, 2a_i + \mathcal{C}_{i-1}, 3ai + \mathcal{C}_{i-1}$; otherwise $\mathcal{C}_i := \mathcal{C}_{i-1}$;
4. $i := i + 1$; return to 2.

For $0 < i \leq n$, the code $\mathcal{C}_i$ are forced to be linear because we take all linear combination of the selected vectors $a_{i1}, \cdots, a_{il}; l \leq i$. The code C_i have "basis" formed selected vectors $a_{i1}, \cdots, a_{il}$. We obtain a nested sequence of linear codes

$$0 = \mathcal{C}_0 \subseteq \mathcal{C}_1 \subseteq \cdots \subseteq \mathcal{C}_n$$

$\mathcal{C}_n$ is the lexicode and we note $\mathcal{C}_n = \mathcal{C}(B, P)$ where B is the ordering and P is the selection property. We obtain the following result.

Theorem 2.1 *([3, Theorem 4]) For any basis B of R^n and any multiplicative selection criterion P, the lexicode $C(B, P)$ is linear and $P[\mathsf{x}]$ holds for each codeword $\mathsf{x} \neq 0$.*

3 A Greedy Algorithm for Bounded GC-content DNA Codes

The elements $\{0, 1, 2, 3\}$ of $\mathbb{Z}_4$ are in one to one correspondence with the nucleotide DNA bases, $\{A, T, C, G\}$, by the map ϕ such that: $0 \rightarrow G$, $2 \rightarrow C$, $3 \rightarrow T$ and $1 \rightarrow A$.

Definition 3.1 *Let $\mathcal{C}$ be a linear code of ${\mathbb{Z}_4}^n$, the GC-content of a codeword $x \in \mathcal{C}$, denoted $GC(\phi(x))$ is the number of occurrence if G and C in $\phi(x)$*

$$GC(\phi(x)) = |\{1 \leq i \leq n;\ \phi(x)_i \in \{G, C\}\}| = w_{GC}(\phi(x))$$

We say that a subset $\mathcal{C}$ of ${\mathbb{Z}_4}^n$, verify the bounded GC-content constraint if there exists a $w \in \mathbb{N}^*$ such that $GC(\phi(x)) \geq w$, $\forall x \in C$.

Proposition 3.2 *The property $p_1[x]$ is true if and only $w_{GC}(x) \geq w$ is a multiplicative property over $\mathbb{Z}_4$*

Table 1: DNA Lexicode over $\mathbb{Z}_4{}^n$ Using the Selection Property $P_2(w_{GC}(x) \geq w)$

n	w	d_H	Basis of $\mathbb{Z}_4$	Basis of $C(B,P)$
8	4	4	Canonical basis	21111000 13210100 32310010
10	6	4	Canonical basis	2111100000 1321010000 3231001000
10	10	1	Canonical basis	2000000000 0200000000 0020000000 0002000000 0000200000 0000020000 0000002000 0000000200 0000000020 0000000002
12	12	1	Canonical basis	200000000000 020000000000 002000000000 000200000000 000020000000 000002000000 000000200000 000000020000 000000002000 000000000200 000000000020 000000000002

3.1 Result From our Computations

In this section we give numerical results of construction of linear codes over $\mathbb{Z}_4$ with bounded GC content by w.
In this case we can eliminate the step of verification for $w_{GC}(\phi(2x)) \geq w$ from the Algorithm A. Because for an $x \in Z_4^n$ if $w_{GC}(\phi(x)) \geq w$ this implies that $w_{GC}(\phi(2x)) \geq w$; which makes the algorithm faster. Some of our codes reach the upper bound (5) of in the paper [4]. Furthermore our codes are linear compared to those given in the paper [6].

4 Edit Distance

We use the edit distance for biological computation specially two types of genetic mutation, the first type is the substitution of nucleotide pair, this type Contains the two model of genetic mutation:
The transition: purine replaced by a purine $(A < - > G)$ or pyrimidine pyrimidine $(T < - > C)$.
The transversion: purine replaced by pyrimidine or inverse (ex: $A < - > C$).
The second type is the modification of part of the drive that are the insertion or deletion.
let $\mathcal{A}$ and $\mathcal{B}$ be a finite of distinct symbols and let $x^t \in \mathcal{A}^t$ denote arbitrary string of the length t over the alphabet $\mathcal{A}$.
A string edit distance is characterized by a triple $< \mathcal{A}, \mathcal{B}, c >$ consists of the finite alphabet $\mathcal{A}$ and $\mathcal{B}$ the primitive function $c : E \rightarrow \mathbb{R}_+$ where $\mathbb{R}_+$ is the set of nonnegative reals,$E = E_s \cup E_d \cup E_i$ is the alphabet of primitive edit operation,$E_s = \mathcal{A} * \mathcal{B}$ is the set of substitution, $E_d = \mathcal{A} * E$ is the set of the deletion and $E_i = E \times \mathcal{B}$ is the set of the insertion.
each such triple $< \mathcal{A}, \mathcal{B}, c >$ induce a distance function $d_c : \mathcal{A}^* \times \mathcal{B}^* \rightarrow \mathbb{R}_+$ that the map of string to a nonnegative value, see article [5].

Table 2: DNA code strand corresponding to the linear code in the second row of Table 1

GGGGGGGGGG	TCTAGGAGGG	GGCCGGCGGG	ACATGGTGGG
ATCAGAGGGG	GAACGAAGGG	TTGTGACGGG	CATGGATGGG
CCGCGCGGGG	AGTTGCAGGG	GCCGGCCGGG	TGAAGCTGGG
TACTGTGGGG	CTAGGTAGGG	AAGCGTCGGG	GTTCGTTGGG
CAAAAGGGGG	ATGCAGAGGG	GATTAGCGGG	TTCGAGTGGG
TGTGAAGGGG	CCCAAAAGGG	AGACAACGGG	GCGTAATGGG
GTATACGGGG	TAGGACAGGG	CTTAACCGGG	AACCACTGGG
ACTGATGGGG	GGCAATAGGG	TCACATCGGG	CGGTATTGGG
GCCCCGGGGG	TGATCGAGGG	CCGGCGCGGG	AGTACGTGGG
AAGTCAGGGG	GTTGCAAGGG	TACACACGGG	CTACCATGGG
CGCGCCGGGG	ACAACCAGGG	CGGCCCCGGG	TCTTCCTGGG
TTGACTGGGG	CATCCTAGGG	ATCTCTCGGG	GAAGCTTGGG
CTTTTGGGGG	AACGTGAGGG	GTAATGCGGG	TAGCGGTGGG
TCAGTAGGGG	CGGATAAGGG	ACTCTACGGG	GGCTTATGGG
GATATCGGGG	TTCCTCAGGG	CAATTCCGGG	ATGGTCTGGG
AGACTTGGGG	GCGTTTAGGG	TGTGTTCGGG	CCCATTTGGG

Table 3: DNA Lexicode over ${\mathbb{Z}_4}^n$ Obtained Using the Selection Property $P_3(d_c(\phi(x),\phi(y)) \leq m)$

n	$\phi(x)$	$d_c(\phi(x),\phi(y)) \leq m$	d_H	w_{GC}	Basis of $\mathbb{Z}_4$	Basis of $C(B,P)$
4	GGGG	1	1	4	Canonical basis	2222 2202 2220 2022
4	GCGC	2	2	4	Canonical basis	2020 0022 0220 2222

Definition 4.1 *The edit distance $d_c(x^t,y^v)$ between two string $x^t \in \mathcal{A}^t$ et $y^v \in \mathcal{B}^v$ is defined recursively*

$$d_c(x^t,y^v) = min \begin{cases} c(x_t,y_v) + d_c(x^{t-1},y^{v-1}) & ; \\ c(x_t,\epsilon) + d_c(x_{t-1},y^v) & ; \\ c(\epsilon,y_v) + d_c(x^t,y^{v-1}) & . \end{cases}$$

Where $d_c(\epsilon,\epsilon) = 0$, where ϵ denotes the empty word of length 0*.*

Proposition 4.2 *The propriety $P_3[x]$ is true and only if $d_c(\phi(x),\phi(y)) \geq w$ is a multiplicative property over $\mathbb{Z}_4$.*

4.1 DNA Lexicode with Edit Distance Criteria

In this section we give numerical results of construction linear code over $\mathbb{Z}_4$ by greedy algorithm with bounded GC-content by w and edit distance $d_c(x,y)$ such that $x \in {\mathbb{Z}_4}^*$ and $y \in {\mathbb{Z}_4}^*$. We must fix the vector x such that $GC(\phi(x)) = w$ and we apply the greedy algorithm with the property $d_c(x,y) \leq m$, such that m is an integer smaller than w.

4.2 Upper bound and edit distance

Define $A_4(n,d)$ to be maximum size of quaternary code withe length n and minimum edit distance d. Define $A_4^{GC}(n,d,n)$ to the maximum size of DNA code with length n, minimum edit distance d and fixed GC weight w, define $A_4^{RC,GC}(n,d,w)$ to the maximum size of DNA code with length n, minimum edit distance d and fixed GC weight w, that satisfy the reverse-complement constraint.

Proposition 4.3 *for $n > 0$, with $0 \leq d \leq n$ and $0 \leq w \leq n$*

$$A_4{}^{GC}(n,d,0) \leq A_2(n,d). \tag{1}$$

$$A_4{}^{GC}(n,d,w) = A_4^{GC}(n,d,n-w). \tag{2}$$

if $w = n/2$ *then*

$$A_4^{GC}(n,d,w) = 4. \tag{3}$$

References

[1] M.A. Bishop, A.G. D'yachkov, A.J. Macula, T.E. Renz, and V.V. Rykov *Free Energy Gap and Statistical Thermodynamic Fidelity of DNA Codes.*Pp.1088-1104. DOI:10.1089/cmb.2007.0083.

[2] Y. M. Chee and S. Ling, Improved Lower Bounds for Constant GC-Content DNA Codes, IEEE Trans. Inform. Theory, 54, Jan 2008.

[3] K. Guenda, T.A. Guilliver and S.A. Sheikholeslam *Lexicodes over Rings,* FirstOnline, Designs Codes Crypt. Feb. 2013.

[4] O.D. King, Bounds for DNA codes with constant GC-content Sep 8, 2003, 05B40.

[5] E.S. Ristad and P.N.Yianilos *Learning String-Edit Distance* Pattern Analysis and machin Intelligence, IEEE Transaction on (volume 20, Issue: 5,1997).

[6] D.H. Smith, N. Aboluion , R. Montemanni and S. Perkins. Linear and nonlinear constructions of DNA codes with Hamming distance d and constant GC-content, Discrete Math. 311, 1207-1219, 2011.

[7] 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, vol 16, 450–456, 1996.

[8] J. Sun,*Bounds on Edit Metric codes with Cominatorial DNA Constraints* Master's thesis, Faculty of mathematic and science, Brock universty, 2009.