Codes With Run-Length and GC-Content Constraints for DNA-Based Data Storage

Wentu Song, Kui Cai, Mu Zhang, and Chau Yuen

***Abstract*—We propose a coding method to transform binary sequences into DNA base sequences (codewords), namely sequences of the symbols A, T, C, and G, that satisfy the following two properties: 1) run-length constraint: the maximum run-length of each symbol in each codeword is at most three and 2) GC-content constraint: the GC-content of each codeword is close to 0.5, say between 0.4 and 0.6. The proposed coding scheme is motivated by the problem of designing codes for DNA-based data storage systems, where the binary digital data is stored in synthetic DNA base sequences. Existing literature either achieve code rates not greater than 1.78 bits per nucleotide or lead to severe error propagation. Our method achieves a rate of 1.9 bits per DNA base with low encoding/decoding complexity and limited error propagation.**

***Index Terms*—DNA data storage, constrained coding, channel coding.**

I. Introduction

In a DNA-based storage system, the binary data is mapped to a large number of DNA sequences (i.e., sequences of symbols A, C, G and T). These DNA sequences are synthesized and stored in a DNA pool. To retrieve the original data, the stored DNA sequences are sequenced and mapped inversely to the binary user data. To combat different types of errors in DNA synthesizing and sequencing, various coding techniques, such as constraint coding and error correction coding, are introduced to DNA-based storage systems [2]–[9].

It has been found that the homopolymer run (i.e., the repetition of the same nucleotide) and GC-content of a DNA sequence (i.e., the ratio of the number of G and C symbols to the length of the sequence) are two major factors affecting the synthesis and sequencing errors [7]. In this letter, we consider the problem of designing high rate coding schemes that encode binary sequences to DNA sequences (codewords) satisfying the following two properties

- **Run-length constraint**: The maximum run-length of each symbol in each codeword is at most three, which is equivalent to the homopolymer run-length constraint required by the DNA storage system;
- **GC-content constraint**: The GC-content of each codeword is close to 0.5, say between 0.4 and 0.6.

Several works in the literature have explored this problem. Grass *et al.* [2] presented a method to encode binary sequences satisfying the run-length constraint and with coding rate 1.78 bits/nt. Some other methods constructing codes with the run-length constraint are presented in [3]–[6], all of which achieve code rate not greater than 1.6 bits/nt.

Based on the sequence replacement technique, Immink, and Cai [8] presented a method for constructing k-constrained q-ary codes, where the run-length of zero is at most k and the code rate is $\frac{n-1}{n}$. A method transforming k-constrained binary sequence into DNA nucleotide strands with homopolymer run at most $\lceil \frac{k}{2} \rceil$ is also proposed in [8].

The DNA fountain method dealing with DNA sequences with both run-length constraint and GC-content constraint was proposed in [7]. Although the rate of the resulted codes is close to the theoretical channel capacity, its iterative decoding process can lead to severe error propagation.

DNA codes with constant GC-content are intensively studied in [10], where theoretical upper and lower bounds on the maximum size of DNA codes with constant GC-content w and minimum Hamming distance d as well as some explicitly construction of codes are presented.

In this letter, we propose a method to map binary sequences into DNA sequences satisfying both the run-length constraint and the GC-content constraint. The basic idea is to concatenate sequences of length n to sequences of length kn by an elaborately constructed adjacency relation, where $3 \leq n \leq 35$ and k is any fixed positive integer. The code rate is $\frac{2n-1}{n}$ bit/nt. Simulated results show that the GC-content of all codewords is between 0.4 and 0.6. For our method, each erroneous nucleotide leads to an erroneous sequence of length $2n$, which is significantly lower than the DNA fountain method.

This letter is organized as follows. In section II, we investigate the properties of quaternary sequences satisfying the run-length constraint. The proposed encoding and decoding methods are presented in Section III. Simulation results are given in Section IV, and the letter is concluded in Section V.

A. Notations and Conventions

We introduce some notations and conventions that will be used throughout this letter.

For mathematical convenience, we use $\mathbb{Z}_4 := \{0, 1, 2, 3\}$ to denote the set of DNA symbols, through the mapping C $\leftrightarrow 0$, T $\leftrightarrow 1$, G $\leftrightarrow 2$ and A $\leftrightarrow 3$. A sequence $C = c_1 c_2 \cdots c_n$ over $\mathbb{Z}_4$ is viewed as an element of $\mathbb{Z}_4^n$. For simplicity, a sequence of length n is also called a length-n sequence. If $C' = c'_1 c'_2 \cdots c'_{n'} \in \mathbb{Z}_4^{n'}$, then $CC' = c_1 c_2 \cdots c_n c'_1 c'_2 \cdots c'_{n'}$ is the concatenation of C and C'. The GC-content of C, denoted by $\text{GC}(C)$, is the ratio of the total number of symbols 0 and 2 in C to the length of C.

For later use, a sequence C over $\mathbb{Z}_4$ is called a *legal* sequence if the maximum run-length of each symbol in C is at most 3, and the value $|\text{GC}(C) - 0.5|$ is called the 0.5-*GC*

Manuscript received May 25, 2018; revised July 14, 2018; accepted August 13, 2018. Date of publication August 22, 2018; date of current version October 8, 2018. This work is supported by SUTD-MIT IDC research grant, Singapore Ministry of Education Academic Research Fund Tier 2 MOE2016-T2-2-054, SUTD-ZJU grant ZJURP1500102, and NSFC 61750110529. The associate editor coordinating the review of this paper and approving it for publication was M. Egan. *(Corresponding author: Wentu Song.)*

The authors are with the Singapore University of Technology and Design, Singapore 487372 (e-mail: wentu_song@sutd.edu.sg; cai_kui@sutd.edu.sg; zhang_mu@sutd.edu.sg; yuenchau@sutd.edu.sg).

Digital Object Identifier 10.1109/LCOMM.2018.2866566

distance of C. Moreover, $\mathbb{Z}_4$ is ordered by $0<1<2<3$ and for any two sequences $C=c_1c_2\cdots c_n$ and $C'=c_1'c_2'\cdots c_n'$ in $\mathbb{Z}_4^n$, we say that C is smaller than C' for the lexicographic order, denoted by $C<_{\text{Lex}}C'$, if $c_i<c_i'$ for the first i where c_i and c_i' differ. Clearly, the set $\mathbb{Z}_4^n$ is totally ordered by $<_{\text{Lex}}$.

If an encoding function maps each length-K binary sequence to a length-N quaternary sequence (codeword), then the code rate is $\frac{K}{N}$ bits per DNA base (bits/nt). Clearly, for any encoding function, it always holds that $\frac{K}{N}\le 2$.

II. Coding With Run-Length Constraint

We first investigate some properties of legal sequences, which will be used in the next section. For our purpose, we introduce some notations as follows.

For any matrix $A=(a_{i,j})_{k\times n}$ such that $a_{i,j}\ge 0$, let

$$\|A\|_{\text{sum-all}} := \sum_{i=1}^{k}\sum_{j=1}^{n} a_{i,j}.$$

For $i\in\{1,2,3\}$, let $M_i=(\lambda_{C,C'})$ be a 64×4^i matrix satisfying: (1) the rows of M_i are indexed by $C\in\mathbb{Z}_4^3$ and the columns of M_i are indexed by $C'\in\mathbb{Z}_4^i$; and (2) $\lambda_{C,C'}=1$ if the concatenation CC' is a legal sequence, and $\lambda_{C,C'}=0$ otherwise. Moreover, let M_0 be the 64×64 identity matrix.

The following proposition gives a method to compute the number of length-n legal sequences.

Proposition 1: For any given positive integer $n\ge 3$, the number of legal sequences of length n, denoted by L_n, is

$$L_n=\|M_3^{q-1}M_i\|_{\text{sum-all}}$$

where $q=\lfloor\frac{n}{3}\rfloor$ ($\lfloor\cdot\rfloor$ is the floor function) and $i=n-3q$.

Proof: Note that we have $n=3q+i$, where $0\le i\le 2$.

First, assume $i=0$, i.e., $3|n$. Consider the directed graph G whose vertex set is $\mathbb{Z}_4^3$ and adjacency matrix is M_3. Clearly, the set of all length-n legal sequences corresponds to the set of all length-$(q-1)$ paths of G. It is well known in graph theory that the number of all length-$(q-1)$ paths of G is the sum of all entries of M_3^{q-1} (e.g., see [11]). So $L_n=\|M_3^{q-1}M_0\|_{\text{sum-all}}$, where M_0 is the identity matrix of order 64.

Now suppose $i\in\{1,2\}$. Then each length-n legal sequence corresponds to a length-$(3q)$ sequence concatenated by a length-i sequence. By a similar way, one can easily check that the number of such sequences is the sum of all entries of $M_3^{q-1}M_i$. That is, $L_n=\|M_3^{q-1}M_i\|_{\text{sum-all}}$.

Hence, for all $n\ge 3$, we always have $L_n=\|M_3^{q-1}M_i\|_{\text{sum-all}}$, which completes the proof. ∎

In the rest of this section, we will assume $n\ge 3$ is an arbitrary fixed integer. For each $C\in\mathbb{Z}_4^3$, let $\mathcal{F}_C$ denote the set of all legal sequences $C'\in\mathbb{Z}_4^n$ such that CC' is a legal sequence.[1] Then we have the following lemma.

Lemma 1: Let L_n be the number of legal sequences of length n. If $\frac{L_n}{4^n}\ge\frac{2}{3}$, then $|\mathcal{F}_C|\ge 2^{2n-1}$ for all $C\in\mathbb{Z}_4^3$.

Proof: Consider the permutation group $\mathscr{S}_4$ of $\mathbb{Z}_4$. For each positive integer ℓ and each $\sigma\in\mathscr{S}_4$, σ induces a permutation of $\mathbb{Z}_4^\ell$ such that $\sigma(c_1c_2\cdots c_\ell)=\sigma(c_1)\sigma(c_2)\cdots\sigma(c_\ell)$ for each

[1]Note that the set $\mathcal{F}_C$ also depends on a fixed positive integer n, that is, $\mathcal{F}_C\subseteq\mathbb{Z}_4^n$. The number n will always be specified by the context.

① # leagal seq
② corollary 1

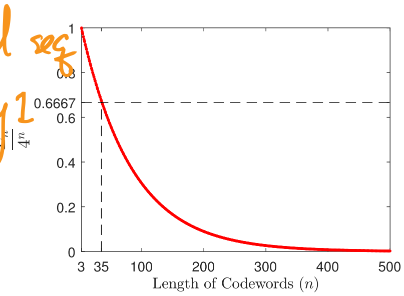

Fig. 1. The ratio of the number of length-n legal sequences (L_n) to the number of all length-n sequences (4^n), where $3\le n\le 500$.

$c_1c_2\cdots c_\ell\in\mathbb{Z}_4^\ell$. Clearly, $c_1c_2\cdots c_\ell$ is a legal sequence if and only if $\sigma(c_1c_2\cdots c_\ell)$ is a legal sequence. So $\mathscr{S}_4$ is also a permutation group of the set of all length-ℓ legal sequences. Moreover, $\mathcal{F}_{\sigma(C)}=\sigma(\mathcal{F}_C)$ for each $C\in\mathbb{Z}_4^3$. So $|\mathcal{F}_C|=|\mathcal{F}_{C'}|$ for any $C,C'\in\mathbb{Z}_4^3$ that are in the same $\mathscr{S}_4$-orbit. By enumeration, we can easily see that $\mathbb{Z}_4^3$ is partitioned into five $\mathscr{S}_4$-orbits, i.e., $\langle 000\rangle$, $\langle 001\rangle$, $\langle 010\rangle$, $\langle 100\rangle$ and $\langle 012\rangle$, where for each $C\in\mathbb{Z}_4^3$, $\langle C\rangle$ denotes the $\mathscr{S}_4$-orbit of C. Hence, it suffices to prove that $|\mathcal{F}_C|\ge 2^{2n-1}$ for $C\in\{000,001,010,100,012\}$.

First consider $\mathcal{F}_{000}$. For $i\in\mathbb{Z}_4$, let $\mathcal{L}_i$ be the set of all legal sequences $c_1c_2\cdots c_n$ such that $c_1=i$. Clearly, $\{\mathcal{L}_0,\mathcal{L}_1,\mathcal{L}_2,\mathcal{L}_3\}$ is a partition of the set of all length-n legal sequences and $\mathcal{F}_{000}=\mathcal{L}_1\cup\mathcal{L}_2\cup\mathcal{L}_3$. Moreover, for each $i\in\{1,2,3\}$, $\mathcal{L}_i=\sigma_i(\mathcal{L}_0)$, where $\sigma_i\in\mathscr{S}_4$ be such that $\sigma_i(0)=i,\sigma_i(i)=0$ and $\sigma_i(j)=j$ for $j\in\mathbb{Z}_4\backslash\{0,i\}$. So we have $|\mathcal{L}_0|=|\mathcal{L}_1|=|\mathcal{L}_2|=|\mathcal{L}_3|$, which implies that $|\mathcal{F}_{000}|=\frac{3}{4}L_n$. Hence, if $\frac{L_n}{4^n}\ge\frac{2}{3}$, then

$$|\mathcal{F}_{000}|=\frac{3}{4}L_n\ge\frac{3}{4}\cdot\frac{2}{3}\cdot 4^n=2^{2n-1}.$$

Further, consider $\mathcal{F}_C$ for $C\in\{001,010,100,012\}$. It is easy to see that $\mathcal{F}_{000}\subseteq\mathcal{F}_{010}$ and $\mathcal{F}_{000}\subseteq\mathcal{F}_{100}$, which implies that $|\mathcal{F}_{010}|\ge|\mathcal{F}_{000}|\ge 2^{2n-1}$ and $|\mathcal{F}_{010}|\ge|\mathcal{F}_{000}|\ge 2^{2n-1}$. Moreover, note that $\mathcal{F}_{111}\subseteq\mathcal{F}_{001}$, $\mathcal{F}_{222}\subseteq\mathcal{F}_{012}$ and $\{111,222\}\subset\langle 000\rangle$. Then we have $|\mathcal{F}_{001}|\ge|\mathcal{F}_{111}|=|\mathcal{F}_{000}|\ge 2^{2n-1}$ and $|\mathcal{F}_{012}|\ge|\mathcal{F}_{222}|=|\mathcal{F}_{000}|\ge 2^{2n-1}$.

By the above discussion, we proved that if $\frac{L_n}{4^n}\ge\frac{2}{3}$, then $|\mathcal{F}_C|\ge|\mathcal{F}_{000}|\ge 2^{2n-1}$ for all $C\in\mathbb{Z}_4^3$. ∎

Corollary 1: Suppose $3\le n\le 35$. Then for all $C\in\mathbb{Z}_4^3$, we have $|\mathcal{F}_C|\ge 2^{2n-1}$.

Proof: By Proposition 1, we can compute $\frac{L_n}{4^n}$ for all positive integers n. In particular, we find that $\frac{L_n}{4^n}\ge\frac{2}{3}$ for $3\le n\le 35$ (see Fig. 1). So by Lemma 1, if $3\le n\le 35$, then $|\mathcal{F}_C|\ge 2^{2n-1}$ for all $C\in\mathbb{Z}_4^3$. ∎

III. Coding With Both Run-Length Constraint and GC Content Constraint

In this section, we present a coding method to map binary sequences of length $k(2n-1)$ to quaternary sequences of length kn that satisfy both the Run-length constraint and the GC-content constraint, where k and n ($n\ge 3$) are two prescribed design parameters. The basic idea is to concatenate k legal sequences, each of length n, say $C_1,C_2,\cdots,C_k$, to obtain a legal sequence $C=C_1C_2\cdots C_k$ of length kn. Since $n\ge 3$, to guarantee that C is legal, it is sufficient that $C'_{i-1}C_i$

is a legal sequence for all $i \in \{2, 3, \cdots, k\}$, where C'_{i-1} is the sequence consisting of the last three symbols of C_{i-1}.

To present our coding method, we first define an order, denoted by $<_{\text{sq}}$, between sequences over $\mathbb{Z}_4$ of the same length. Specifically, for any two different sequences C and C' of the same length, $C <_{\text{sq}} C'$ if and only if one of the following two conditions holds: (1) the 0.5-GC distance of C is smaller than the 0.5-GC distance of C'; (2) the 0.5-GC distance of C equals the 0.5-GC distance of C' and $C <_{\text{Lex}} C'$. Clearly, for each positive integer n, the set $\mathbb{Z}_4^n$ is totally ordered by $<_{\text{sq}}$.

In the rest of this section, we still suppose n is a fixed integer satisfying $3 \leq n \leq 35$. For each $C \in \mathbb{Z}_4^3$, by Corollary 1, we have $|\mathcal{F}_C| \geq 2^{2n-1}$, where $\mathcal{F}_C$ denotes the set of all legal sequences $C' \in \mathbb{Z}_4^n$ such that CC' is a legal sequence. Let $\mathcal{G}_C \subseteq \mathcal{F}_C$ consisting of the smallest 2^{2n-1} elements of $\mathcal{F}_C$ with respect to the order $<_{\text{sq}}$. Moreover, let

$$\mathcal{G} = \bigcup_{C \in \mathbb{Z}_4^3} \mathcal{G}_C$$

and $\mathcal{G}_{\text{O}}$ be the smallest 2^{2n-1} elements of $\mathcal{G}$ with respect to the order $<_{\text{sq}}$. Clearly, $|\mathcal{G}| \geq 2^{2n-1}$. Sequences in $\mathcal{G}$ will serve as the basic units used to construct legal sequences of length kn. By the construction of $\mathcal{G}$, the 0.5-GC distance of the sequences in $\mathcal{G}$ is as small as possible, hence, the GC-content is as close to 0.5 as possible.

For each $C \in \mathbb{Z}_4^3 \cup \{\text{O}\}$, let the elements of $\mathcal{G}_C$ be listed in ascending order with respect to $<_{\text{sq}}$ and let

$$\xi_C : \mathbb{Z}_2^{2n-1} \to \mathcal{G}_C$$

be such that for each $B \in \mathbb{Z}_2^{2n-1}$, viewing B as a base-2 integer, then $\xi_C(B)$ is the Bth element of $\mathcal{G}_C$. Moreover, let

$$\xi_C^{-1} : \mathcal{G}_C \to \mathbb{Z}_2^{2n-1}$$

be the inverse of ξ_C. Since $|\mathcal{G}_C| = 2^{2n-1} = |\mathbb{Z}_2^{2n-1}|$, so ξ_C and ξ_C^{-1} can always be constructed. Finally, let

$$\eta : \mathbb{Z}_4^n \to \mathbb{Z}_4^3$$

such that for each $C \in \mathbb{Z}_4^n$, $\eta(C)$ is the sequence consisting of the last three symbols of C.

A. Encoding Scheme

The encoding function

$$f_e : \mathbb{Z}_2^{k(2n-1)} \to \mathbb{Z}_4^{kn}$$

is defined as follows. Let $B = B_1 B_2 \cdots B_k \in \mathbb{Z}_2^{k(2n-1)}$ be any given binary sequence of length $k(2n-1)$, where each B_i is a binary sequence of length $2n-1$ and is viewed as an element of $\mathbb{Z}_2^{2n-1}$. Then $f_e(B) = C_1 C_2 \cdots C_k$ such that

$$C_1 = \xi_{\text{O}}(B_1)$$

and for $2 \leq i \leq k$,

$$C_i = \xi_{C'_{i-1}}(B_i)$$

where

$$C'_{i-1} = \eta(C_{i-1}).$$

By the construction, $C_i = \xi_{C'_{i-1}}(B_i) \in \mathcal{G}_{C'_{i-1}} = \mathcal{G}_{\eta(C_{i-1})}$ for each $i \in \{2, \cdots, k\}$. So $\eta(C_{i-1})C_i$ is a legal sequence, where $\eta(C_{i-1})$ consists of the last three symbols of C_{i-1}. Moreover, since each C_i, $i = 1, 2, \cdots, k$, is a legal sequence, so $f_e(B) = C_1 C_2 \cdots C_k$ is a legal sequence.

As an illustrative example, let $n = 4$, $k = 3$. First, for each $C \in \mathbb{Z}_4^3 \cup \{O\}$, we can construct a subset $\mathcal{G}_C \subseteq \mathcal{F}_C \subseteq \mathbb{Z}_4^n$ and a map $\xi_C : \mathbb{Z}_2^{2n-1} \to \mathcal{G}_C$. Limited by the space, we omitted the details of $\mathcal{G}_C$ and ξ_C. Let $B = B_1 B_2 B_3 = 001000010110110010100$ be a binary sequence, where $B_1 = 0010000$, $B_2 = 1011011$ and $B_3 = 0010100$. Then we can obtain $C_1 = \xi_O(B_1) = 0012$, $C_2 = \xi_{C'_1}(B_2) = 0122$ and $C_3 = \xi_{C'_2}(B_2) = 0122$, where $C'_1 = \eta(C_1) = 012$ and $C'_2 = \eta(C_2) = 122$. Finally, $f_e(B) = C_1 C_2 C_3 = 001201220321$.

We now estimate the probability distribution of the GC-content of the codewords $f_e(B) = C_1 C_2 \cdots C_k$. Let X and X_i be the random variable representing the GC-content of $f_e(B)$ and C_i, $i = 1, 2, \cdots, k$, respectively. Since $C_1, C_2, \cdots, C_k$ have the same length n, then we have

$$X = \frac{X_1 + X_2 + \cdots + X_k}{k}.$$

To simplify discussion, we assume that $X_1, X_2, \cdots, X_k$ are i.i.d. random variables.[2] Since each C_i has length n, then X_i takes values in the set $\{c_j = \frac{j}{n}; j = 0, 1, \cdots, n\}$ and

$$\Pr[X_i = c_j] = \sum_{C' \in \mathcal{G}, \text{GC}(C') = c_j} \Pr[C_i = C']. \tag{1}$$

For each $C' \in \mathcal{G}$, let $N_{C'}$ denote the number of $C'' \in \mathcal{G}$ such that $C' \in \mathcal{G}_{\eta(C'')}$. By our encoding scheme, the existence of a $C'' \in \mathcal{G}$ with $C' \in \mathcal{G}_{\eta(C'')}$ gives a chance of $C_i = C'$ (i.e., a chance of $C_{i-1}C_i = C''C'$). So we can approximate the probability as

$$\Pr[C_i = C'] = \frac{N_{C'}}{\sum_{C \in \mathcal{G}} N_C}. \tag{2}$$

From (1) and (2), we have

$$\Pr[X_i = c_j] = \frac{1}{\sum_{C \in \mathcal{G}} N_C} \left(\sum_{C' \in \mathcal{G}: \text{GC}(C') = c_j} N_{C'} \right). \tag{3}$$

Let

$$\Gamma = \left\{ (r_0, r_1, \cdots, r_n); \sum_{j=0}^{n} r_j = k, r_j \geq 0 \right\}$$

and for each real number $d \geq 0$, let

$$\Gamma_d = \left\{ (r_0, r_1, \cdots, r_n) \in \Gamma; \frac{1}{k} \sum_{j=0}^{n} r_j c_j = d_\ell \right\}.$$

Since Γ is a finite set, there are only finite number of values of d, denoted by $d_1, d_2, \cdots, d_T$, such that $\Gamma_d \neq \emptyset$. Denote

[2]Intuitively, a more accurate model for $X_1, X_2, \cdots, X_k$ is Markov Chain. However, the Markov Chain model has a higher computational complexity. On the other hand, our simulation results shows that the i.i.d random series model can well approximate the GC-content distribution (see Section IV).

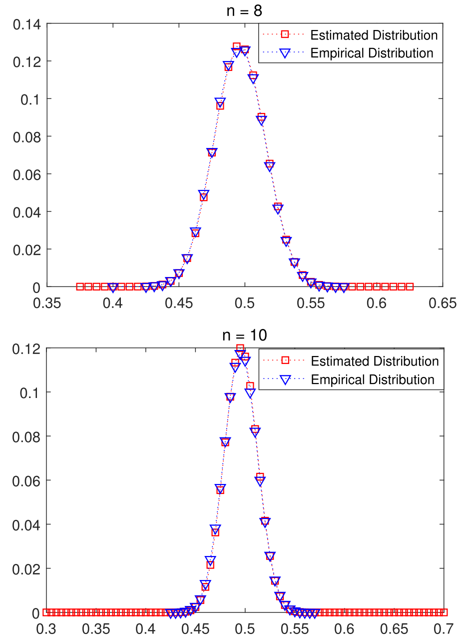

Fig. 2. The estimated distribution and empirical distribution of the codewords' GC-content for $n = 8$ and $n = 10$.

$\Pr[X_i = c_j] = \theta_j$. Then for each $\ell \in \{1, 2, \cdots, T\}$, we have

$$\begin{aligned}\Pr[X = d_\ell] &= \sum_{(r_0, r_1, \cdots, r_n) \in \Gamma_{d_\ell}} \mathrm{Mu}((r_0, r_1, \cdots, r_n) | k, \boldsymbol{\theta}) \\ &= \sum_{(r_0, r_1, \cdots, r_n) \in \Gamma_{d_\ell}} \frac{k!}{r_0! \; r_1! \; \cdots \; r_n!} \prod_{j=0}^{n} \theta_j^{r_j} \qquad (4)\end{aligned}$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_n)$ and $\mathrm{Mu}((r_0, r_1, \cdots, r_n)|k, \boldsymbol{\theta})$ is the multinomial distribution with parameters k and $\boldsymbol{\theta}$ [12, Ch. 2]. Simulation results show that the GC-content of the codewords are close to 0.5 as expected (see Section IV).

Note that the code rate of our scheme is $\frac{2n-1}{n}$ bits/nt, which increases with the increase of n. On the other hand, the main cost of the encoding process is to construct the set $\mathcal{G}_C$ for each $C \in \mathbb{Z}_4^3$, which is time consuming for large n (say for $n \geq 20$). In practice, n should be properly chosen to balance the code rate and encoding complexity. Our simulation shows that for $n = 10$, the collection $\{\mathcal{G}_C; C \in \mathbb{Z}_4^3\}$ can be quickly constructed and the encoding can be realized very efficiently. In the mean while, for $n = 10$, the code rate is $\frac{19}{10} = 1.9$ bits/nt, which is higher than most of the existing coding scheme.

B. Decoding Algorithm

The decoding function

$$f_d : \mathbb{Z}_4^{kn} \to \mathbb{Z}_2^{k(2n-1)}$$

is defined as follows. Let $C = C_1 C_2 \cdots C_k \in \mathbb{Z}_4^{kn}$ be a received sequence, where each C_i is a legal sequence of length n. Then $f_d(C) = B_1 B_2 \cdots B_k$ such that

$$B_1 = \xi_O^{-1}(C_1)$$

and for $2 \leq i \leq k$,

$$B_i = \xi_{C'_{i-1}}^{-1}(C_i)$$

where $C'_{i-1} = \eta(C_{i-1})$. It is easy to verify that for any $B \in \mathbb{Z}_2^{k(2n-1)}$, if $C = f_e(B)$, then $B = f_d(C)$.

To decode $C_1 C_2 \cdots C_k$, one need to compute $\xi_O^{-1}(C_1)$ and $\xi_{C'_{i-1}}^{-1}(C_i)$ for each $i \in \{2, \cdots, k\}$, that is, to find the location of each C_i in the ordered set $\mathcal{G}_{C'_{i-1}}$. This can be done by the binary search algorithm in time $O(\log_2 |\mathcal{G}_{C'_{i-1}}|) = O(n)$.

IV. Simulation Results

We set $k = 20$ and randomly choose 10^5 binary sequences of length $k(2n-1)$ for each $n \in \{5, 6, \cdots, 12\}$. The simulation results show that the GC-content of all coded sequences is between 0.4 and 0.6. Limited by the space, we give the estimated distribution and empirical distribution of the codewords' GC-content only for $n \in \{8, 10\}$, see Fig. 2.

V. Conclusions

We propose a method to encode binary sequences to quaternary sequences (codewords) that satisfy both the run-length constraint and the GC-content constraint. Our method achieves rate of $\frac{2n-1}{n}$ bits/nt and simulation results show that the GC-contents are between 0.4 and 0.6, where $3 \leq n \leq 35$. For $n = 10$, our method achieves code rate 1.90 bits/nt and has low encoding/decoding complexity.

Our method can also be incorporated with the concept of Hamming distance and edit distance to enhance the reliability for practical purpose. We leave it in our future work.

References

[1] J. Davis, "Microvenus," *Art J.*, vol. 55, no. 1, pp. 70–74, 1996, doi: 10.2307/777811.

[2] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angew. Chem. Int. Ed.*, vol. 54, no. 8, pp. 2552–2555, 2015.

[3] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, p. 1628, 2012.

[4] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-based archival storage system," in *Proc. 21st Int. Conf. Architectural Support Program. Lang. Operating Syst.*, 2016, pp. 637–649.

[5] N. Goldman *et al.*, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, pp. 77–80, Jan. 2013.

[6] M. Blawat *et al.*, "Forward error correction for DNA data storage," *Procedia Comput. Sci.*, vol. 80, pp. 1011–1022, Jan. 2016.

[7] Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, 2017.

[8] K. A. S. Immink and K. Cai, "Design of capacity-approaching constrained codes for DNA-based storage systems," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 224–227, Feb. 2018.

[9] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Sci. Rep.*, vol. 5, Sep. 2015, Art. no. 14138.

[10] O. D. King, "Bounds for DNA codes with constant GC-content," *Electron. J. Combinatorics*, vol. 10, no. 1, p. 33, 2003.

[11] D. Cartwright and T. C. Gleason, "The number of paths and cycles in a digraph," *Psychometrika*, vol. 31, no. 2, pp. 179–199, 1966.

[12] K. P. Murphy, *Machine Learning—A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.