Constrained Channel Capacity for DNA-Based Data Storage Systems

Kaixin Fan, Huaming Wu, *Senior Member, IEEE*, and Zihui Yan

***Abstract*— Deoxyribonucleic acid (DNA)-based data storage has grown rapidly due to its advantages with the increase in infrequently large amounts of data. However, when the maximum homopolymer runlength (RLL) of the DNA strand is large and the GC-content is either too high or too low, the DNA synthesis and sequencing processes are prone to substitution, deletion and insertion errors. To reduce errors in DNA synthesis and sequencing, we require that the DNA storage channel satisfies both k-RLL and strong-(l,δ)-locally-GC-balanced constraints, where the former refers to the maximum homopolymer runlength in each sequence is at most k, and the latter refers to the number of G and C of every length-($l' \geq l$) subsequence is bounded between $[\frac{l'}{2} - \delta, \frac{l'}{2} + \delta]$. This constrained channel allows DNA data storage system to be less prone to errors during synthesis and sequencing and improves the success rate of Polymerase Chain Reaction (PCR) amplification. We propose a method to calculate the channel capacity. In particular, we provide a relationship between the 4-ary constrained channel capacity and the 2-ary constrained channel capacity, which makes it simpler to calculate the 4-ary constrained channel capacity.**

***Index Terms*— DNA-based storage systems, constrained channels, channel capacity.**

I. Introduction

In THE era of big data, the demand for cold data storage has increased, and the ever-increasing data of humans has brought new pressure on storage methods. Traditional storage methods such as magnetic disks, optical disks and hard disks are no longer sufficient. Deoxyribonucleic acid (DNA)-based data storage has the characteristics of high storage density, extremely long storage life, low maintenance costs and convenient data backup. All these advantages make DNA storage very promising. DNA in nature is a biological macromolecule that carries genetic information in an organism. It consists of four deoxynucleotides carrying different bases, namely, adenine (A), cytosine (C), guanine (G), and thymine (T). The process of DNA data storage includes information encoding, DNA synthesis, DNA storage, DNA sequencing and information decoding. The first thing DNA data storage does is encode binary information as a sequence of four different nucleotides arranged in a DNA molecule.

Data information is mainly stored in the form of short DNA fragments due to the limitation of the storage process in DNA data storage. DNA strands need to meet some basic constraints to reduce errors during synthesis and sequencing. Recently, many experimental works based on DNA storage have achieved success [1], [2], [3], [4], [5], [6], [7]. Especially, in 2019, Shankland et al. [6] encoded all 16 GB of text in the English version of Wikipedia into synthetic DNA. In 2021, Roquet et al. [7] developed a custom DNA data writer capable of writing data to DNA at 18 Mbps. On the one hand, Ross et al. [8] experimentally found that substitution and deletion errors increase significantly when the length of homopolymer runs (i.e. repeats of the same nucleotide) in the nucleotide chain exceeds 6. On the other hand, the global GC-content (i.e. the percentage of nucleotides containing G and C in the nucleotide chain in the length of the entire nucleotide chain) is too high or too low, which will easily make the nucleotide chain synthesis and sequencing errors occurred [8], [9]. Thus, we should reduce the homopolymer runlength and keep the global GC-content within a certain range when designing DNA strands. That is, k-RLL constraint (maximum homopolymer runlength is no more than k) and global GC-content constraint are imposed on the sequences. In addition, Benita et al. [10] found that regionalized GC-content, which is more restrictive than global GC-content, is a good predictor of Polymerase Chain Reaction (PCR) success and is more sensitive than the previously described parameters in predicting PCR outcome. Regionalized GC-content constraint can improve the success rate of PCR amplification. So strong-(l, δ)-locally-GC-balanced constraint (for windows with length greater than or equal to l, the GC-content is within a certain range) is particularly important, and we should consider it.

For asymptotically large codeword length n, the maximum number of (binary) user bits that can be stored per q-ary symbol, called (information) capacity, denoted by $\mathbb{C}$. Without any constraint, a nucleotide can store up to 2 bits. As more and more constraints are built into the encoded message, the average amount of information contained in each nucleotide decreases, i.e., the channel capacity decreases. In DNA storage, it is an upper bound on the ratio of the number of random data bits in a sequence to the number of nucleotides required to encode the average of all possible sequences. For all constrained codes, the channel capacity is less than 2. Because it's a theoretical upper bound, it's usually not obtained in the implementation of a particular code. In order to provide a reference for subsequent coding methods that satisfy both k-RLL constraint and strong-(l, δ)-locally-GC-balanced constraint, we calculate its channel capacity.

On the k-RLL constraint, Erlich and Zielinski [2] used the probability to calculate the size of the set of valid codewords under this constraint. Its capacity formula shows that it is only related to the maximum homopolymer runlength and not to the sequence length. Immink and Cai [11] used a generator function to enumerate q-ary sequences of length n that satisfy the k-RLL constraint, and finally obtained the value of the capacity through Shannon's approach [12]. On the strong-(l, δ)-locally-balanced constraint, Gabrys et al. [13] proposed the concept of strong-(l, δ)-locally-balanced constraint, whose capacity is independent of l and is the same as the capacity

Manuscript received 28 August 2022; accepted 2 October 2022. Date of publication 5 October 2022; date of current version 9 January 2023. This work is supported by the National Key R&D Program of China (2020YFA0712100). The associate editor coordinating the review of this letter and approving it for publication was Y. Deng. *(Corresponding author: Huaming Wu.)*

The authors are with the Center for Applied Mathematics, Tianjin University, Tianjin 300072, China (e-mail: kxfan@tju.edu.cn; whming@tju.edu.cn; yanzh@tju.edu.cn).

Digital Object Identifier 10.1109/LCOMM.2022.3212200

of the $(2\delta+1)$-RDS constraint. However, they only discuss the 2-ary case and neglect the important limitation of k-RLL constraint. Nguyen et al. [14] only focused on the maximum homopolymer runlength constraint and ignored the regionalized GC-content constraint. Erlich et al. [2] only considered the global GC-content constraint, which plays a negligible role in reducing the information content of each nucleotide under certain conditions. Unfortunately, no one has yet considered imposing both constraints on sequences at the same time.

To fill this gap, we focus on providing a DNA storage channel model that incorporates both k-RLL constraint and strong-(l,δ)-locally-GC-balanced constraint, and calculate its capacity to guide the DNA data coding scheme. The main contributions of this work are three-fold:

- We combine the k-RLL constraint and strong-(l,δ)-locally-GC-balanced constraint to obtain a novel constrained channel model for DNA storage. To the best of our knowledge, this is the first effort to combine these two constraints simultaneously. This constrained channel allows DNA data storage system to be less prone to errors during synthesis and sequencing and improves the success rate of PCR amplification.
- A method for calculating the capacity of the above-mentioned multiple constrained channel is provided. In particular, we provide a relationship between the 4-ary constrained channel capacity and the 2-ary constrained channel capacity. This work is the first to obtain accurate values for k-RLL constrained and strong-(l,δ)-locally-GC-balanced constrained channel capacity.
- The capacity $\mathbb{C}$ is the upper bound on the rate at which any encoding scheme can convert arbitrary data into a sequence with given constraints. By calculating the capacity of the DNA storage channel that satisfies these two limits, it is possible to determine whether this scenario is suitable for DNA storage, thus giving some motivation to subsequent coding and error correction algorithms.

II. Definitions and Preliminaries

For the convenience of operation, there is a one-to-one correspondence between $\{A,T,C,G\}$ and $\Sigma_4=\{0,1,2,3\}$, i.e. $A\leftrightarrow 0, T\leftrightarrow 1, C\leftrightarrow 2, G\leftrightarrow 3$. The relevant definitions are as follows:

Definition 1 (Maximum Homopolymer Runlength Constraint (or k-Runlength Limited Constraint, or k-RLL Constraint)): Let $\boldsymbol{\sigma}=(\sigma_1,\sigma_2,\ldots,\sigma_n)\in\Sigma_4^n=\{0,1,2,3\}^n$. Given $k>0$, we say that $\boldsymbol{\sigma}$ is k-runlength limited if the run of consecutive nucleotide of Σ_4 is at most k.

Definition 2 (Strong-(l,δ)-Locally-GC-Balanced Constraint): Let l be an even positive integer and δ be a nonnegative integer. A DNA sequence $\boldsymbol{\sigma}\in\Sigma_4^n$ is said to satisfy the strong-(l,δ)-locally-GC-balanced constraint (or the sequence is strong-(l,δ)-locally-GC-balanced) if for all even $l'\geq l$ and $1\leq i\leq n-l'+1$, the number of 3(G) and 2(C) in the window $\boldsymbol{\sigma}[i;l']=(\sigma_i,\sigma_{i+1},\ldots,\sigma_{i+l'-1})$ is in $\left[\frac{l'}{2}-\delta,\frac{l'}{2}+\delta\right]$, i.e.

$$\frac{l'}{2}-\delta\leq w(\boldsymbol{\sigma}[i;l'])\leq\frac{l'}{2}+\delta. \tag{1}$$

TABLE I
A Quaternary Model of DNA Data Storage Coding

binary data	00	01	10	11
base	A	T	C	G
quaternary data	0	1	2	3

We adopt a quaternary model of DNA data storage coding table based on the base characteristics of DNA [15], as shown in Table I. Therefore, binary sequences and DNA sequences can be freely converted by the rule that two bits correspond to one base, which can be expressed as

$$\sigma_i=2x_{2i-1}+x_{2i},\ \ 1\leq i\leq n,\ \ \forall\boldsymbol{\sigma}\in\Sigma_4^n,\boldsymbol{x}\in\Sigma_2^{2n}. \tag{2}$$

And the binary sequence $\boldsymbol{x}\in\Sigma_2^{2n}$ can be divided into odd sequence $\boldsymbol{x_o}=(x_1,x_3,\ldots,x_{2n-1})$ and even sequence $\boldsymbol{x_e}=(x_2,x_4,\ldots,x_{2n})$.

Definition 3 (2-qry k-Runlength Limited Constraint (or 2-qry k-RLL Constraint)): Let $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)\in\Sigma_2^n=\{0,1\}^n$. Given $k>0$, we say that $\boldsymbol{x}$ is 2-qry k-runlength limited if the run of consecutive symbol of Σ_2 is at most k.

For a word $\boldsymbol{x}\in\Sigma_2^n=\{0,1\}^n$, $wt(\boldsymbol{x})$ represents its Hamming weight.

Definition 4 (Strong-(l,δ)-Locally-Balanced Constraint): Let l be an even positive integer and δ be a nonnegative integer. A word $\boldsymbol{x}\in\Sigma_2^n$ is said to satisfy the strong-(l,δ)-locally-balanced constraint(or the word is strong-(l,δ)-locally-balanced) if for all even $l'\geq l$ and $1\leq i\leq n-l'+1$, it holds that

$$\frac{l'}{2}-\delta\leq wt(\boldsymbol{x}[i;l'])\leq\frac{l'}{2}+\delta. \tag{3}$$

Shannon [12] defined the capacity $\mathbb{C}$ of a constrained channel by

$$\mathbb{C}=\lim_{m\to\infty}\frac{1}{m}\log_2 N(m), \tag{4}$$

where $N(m)$ denotes the number of admissible sequences of length m.

Definition 5 (Various Capacity Calculations): Let $\mathbb{S}_q^{rll}(n,k)$ be the set of all q-ary k-runlength limited words of length n. Let $\mathbb{S}_q^{slb}(n,\geq l,\delta)$ be the set of all strong-(l,δ)-locally(-GC)-balanced words of length n. The capacity of these sets are

$$\mathbb{C}_q(k)=\limsup_{n\to\infty}\frac{\log\left(\left|\mathbb{S}_q^{rll}(n,k)\right|\right)}{n}, \tag{5}$$

$$\mathbb{C}_q(\geq\ell,\delta)=\limsup_{n\to\infty}\frac{\log\left(\left|\mathbb{S}_q^{slb}(n,\geq l,\delta)\right|\right)}{n}. \tag{6}$$

Moreover, the capacity of the DNA storage channel that satisfies k-RLL constraint and strong-(l,δ)-locally-GC-balanced constraints:

$$\mathbb{C}_{DNA}(k,\geq l,\delta)=\limsup_{n\to\infty}\frac{\log\left(\left|\mathbb{S}_4^{rll}(n,k)\cap\mathbb{S}_4^{sGC}(n,\geq l,\delta)\right|\right)}{n}.$$

Regarding strong-(l,δ)-locally-balanced constraint, we need the following definitions.

TABLE II
SYMBOLS AND DEFINITIONS

Symbol	Definition
$w(\boldsymbol{\sigma})$	The weight of the DNA strand
$\boldsymbol{\sigma}[i;l]$	A consecutive subsequence of length l starting at i-th coordinate
$\boldsymbol{x_o}$	The odd sequence of $\boldsymbol{x} \in \Sigma_2^{2n}$ and the length is n
$\boldsymbol{x_e}$	The even sequence of $\boldsymbol{x} \in \Sigma_2^{2n}$ and the length is n
$\mathbb{S}_q^{rll}(n,k)$	The set of all q-ary k-runlength limited words of length n
$\mathbb{S}_q^{slb}(n,\geq l,\delta)$	The set of all strong-(l,δ)-locally(-GC)-balanced words of length n
$\mathbb{C}_q(k)$	Capacity of $\mathbb{S}_q^{rll}(n,k)$, set of all q-ary k-runlength limited words
$\mathbb{C}_q(\geq \ell,\delta)$	Capacity of $\mathbb{S}_q^{slb}(n,\geq l,\delta)$, set of all strong-($l$,$\delta$)-locally-balanced words
$\mathbb{C}_{DNA}(k,\geq l,\delta)$	Capacity of channel satisfying k-RLL & strong-(l,δ)-locally-GC-balanced
$\mathbb{S}_2^{RDS}(n,M)$	The set of all M-RDS words of length n
$\mathbb{C}_{RDS}(M)$	Capacity of $S_2^{RDS}(n,M)$, set of all M-RDS words of length n
$Diff(\boldsymbol{x})$	The differential of $\boldsymbol{x} \in \Sigma_2^n$
$Diff^{-1}(\boldsymbol{y})$	The inverse of $Diff(\boldsymbol{x})$
$\mathbb{S}_2^o(n,k,\geq l,\delta)$	Set of 2-ary sequences satisfying k-RLL & strong-(l, δ)-locally-balanced
$\mathbb{S}_2^{cRDS}$	Set of 2-ary sequences satisfying r-constrained and M-RDS
$\mathbb{C}_2^o(k,\geq l,\delta)$	Capacity of the set of odd sequences satisfying constraints
$\mathbb{C}_2^{cRDS}(k-1,M)$	Capacity of 2-ary $(k-1)$-constrained & M-RDS constrained channel

Definition 6 (Running Digital Sum Sequence): For $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \Sigma_2^n$*, its running digital sum sequence* $RDS(\boldsymbol{x}) = (s_0, s_1, \ldots, s_n) \in \mathbb{Z}^{n+1}$ *is defined as follows:* $s_0 = 0$*, and for* $1 \leq i \leq n$*,*

$$s_i = \Sigma_{j=1}^{i}(-1)^{1-x_j} = 2wt((x_1, \ldots, x_i)) - i.$$

The disbalance of $\boldsymbol{x}$ *is* $dis(\boldsymbol{x}) = \max_{0\leq i\leq n} s_i - \min_{0\leq i\leq n} s_i$.

Definition 7 ($M-RDS$ *Words): For a given positive integer* M*, a word* $\boldsymbol{x} \in \Sigma_2^n$ *is called the* $M - RDS$ *word if* $dis(\boldsymbol{x}) \leq M$*. And* $\mathbb{S}_2^{RDS}(n,M)$ *is the set of all* $M - RDS$ *words of length* n *and the capacity of this constraint is*

$$\mathbb{C}_{RDS}(M) = \limsup_{n\to\infty} \frac{\log\left(\left|\mathbb{S}_2^{RDS}(n,M)\right|\right)}{n}. \tag{7}$$

According to [13], the principle of set inclusion have shown that every $(2\delta+1)$-RDS sequence is a strong-(l,δ)-locally-balanced sequence, for all $\delta > 0$, $\ell \geq 4$. And the set $\mathbb{S}_2^{RDS}(n, 2\delta+1)$ is asymptotically optimal for strong-(l,δ)-locally-balanced constraint. That is,

$$\mathbb{S}_2^{RDS}(n, 2\delta+1) \subseteq \mathbb{S}_2^{slb}(n, \geq l, \delta), \tag{8}$$

$$\mathbb{C}_{RDS}(2\delta+1) = \mathbb{C}_2(\geq \ell, \delta). \tag{9}$$

Regarding 2-ary k-runlenth limited constraint, we need the following definition.

Definition 8 (2-ary r*-Constrained): The binary sequences* $\boldsymbol{y}$ *of length* n *is 2-ary* r*-constrained if the number of consecutive "0"s of* $\boldsymbol{y}$ *is no more than* r*.*

Unlike 2-qry k-runlength limited constraint, there is no limit to the number of "1".

Definition 9 (Differential of the Binary Sequence): For a binary sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \Sigma_2^n$*, the binary sequence* $\boldsymbol{y} = Diff(\boldsymbol{x}) = (y_1, y_2, \ldots, y_n) \in \Sigma_2^n$ *is the differential of* $\boldsymbol{x}$*, where* $y_1 = x_1$ *and* $y_i = x_i - x_{i-1} (\mathrm{mod} 2)$ *for* $2 \leq i \leq n$ *[14].*

From the above definition, if we know $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \Sigma_2^n$, we can uniquely determine $\boldsymbol{x} = Diff^{-1}(\boldsymbol{y})$ by

$$x_i = \Sigma_{j=1}^{i} y_j (\mathrm{mod} 2), 1 \leq i \leq n. \tag{10}$$

For the convenience of the reader, the relevant symbols involved in this letter are summarized in Table II.

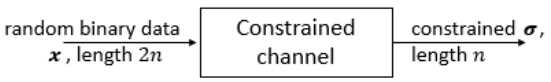

Fig. 1. The constrained channel converts $\boldsymbol{x_e}$ and constrained $\boldsymbol{x_o}$ to constrained $\boldsymbol{\sigma}$ based on Eq. (2).

III. CAPACITY OF k-RLL AND STRONG-(l,δ)-LOCALLY-GC-BALANCED CONSTRAINED CHANNEL

In this section, we will calculate the capacity of the 2-ary k-RLL and strong-(l,δ)-locally-GC-balanced constrained channel for all values of k, l and δ. We convert the quaternary channel capacity calculation into a binary calculation through a proposition. This makes calculations simple. As shown in Fig. 1, the model diagrams give the reader a clearer understanding of Proposition 1.

proposition 1: The 4-ary sequence $\boldsymbol{\sigma} \in \Sigma_4^n$ satisfies both the k-runlength limited constraint and strong-(l,δ)-locally-GC-balanced constraint if and only if its corresponding 2-ary odd sequence $\boldsymbol{x_o} \in \Sigma_2^n$ satisfies both the 2-ary k-runlength limited constraint and strong-(l,δ)-locally-balanced constraint.

Proof: On the one hand, it can be easily seen from Table I that when the symbol in the odd sequence $\boldsymbol{x_o}$ is "1", the corresponding symbol of the 4-ary sequence $\boldsymbol{\sigma}$ obtained by Eq. (2) is "2" or "3". Therefore, when the odd sequence $\boldsymbol{x_o}$ satisfies the strong-(l,δ)-locally-balanced constraint if and only the quaternary sequence $\boldsymbol{\sigma}$ satisfies the strong-(l,δ)-locally-GC-balanced constraint. On the other hand, we need to prove that the 4-ary sequence $\boldsymbol{\sigma}$ satisfies the k-runlength limited constraint if and only if the binary odd sequence $\boldsymbol{x_o}$ satisfies the 2-ary k-runlength limited constraint.

A. Adequacy

Suppose that $\boldsymbol{x_o}$ is 2-ary k-runlength limited, then the maximum run of consecutive identical symbols for $\boldsymbol{x_o}$ is k. Without loss of generality, it is assumed that the odd sequence $\boldsymbol{x_o}$ has k consecutive repeated "1"s. According to Eq. (2), the consecutive k symbols of the corresponding 4-ary sequence $\boldsymbol{\sigma}$ ar"2" or "3". For example, let $n = 5, k = 3$, and the 2-ary odd sequence is $\boldsymbol{x}_o = 01110$. According to Eq. (2), we have $\boldsymbol{\sigma}$ =0(or 1)2(or 3)2(or 3)2(or 3)0(or 1). Thus the number of consecutive "2" or "3" of this $\boldsymbol{\sigma}$ will not exceed k. Then the number of consecutive repetitions of the same symbol in the $\boldsymbol{\sigma}$ does not exceed k, which is k-runlength limited.

B. Necessity

proof by contradiction. Assuming that the 2-ary sequence $\boldsymbol{x_o}$ is not 2-ary k-runlength limited, then the number of consecutive repetitions of the same symbol in $\boldsymbol{x_o}$ is greater than k. Without loss of generality, assuming that there are $k+1$ consecutive "1"s in $\boldsymbol{x_o}$. If the corresponding $k+1$ symbols in $\boldsymbol{x_e}$ of $\boldsymbol{\sigma}$ are all "0"s or all "1"s, then according to Eq. (2), the corresponding $k+1$ symbols of $\boldsymbol{\sigma}$ are all "2"s or all "3"s. This is inconsistent with $\boldsymbol{\sigma}$ being k-runlength limited. A contradiction arises, so Necessity is proven. □

According to the above proposition, the capacity of the constrained DNA channel can be calculated by $\mathbb{C}_{DNA}(k,\geq l,\delta) = 1 + \mathbb{C}_2^o(k,\geq l,\delta)$, where $\mathbb{C}_2^o(k,\geq l,\delta)$ is

the capacity of the set of odd sequences. Therefore, we only need to calculate the capacity of the channel where the constrained 2-ary odd sequences $\boldsymbol{x_o}$ are located. In the following parts, we will calculate $\mathbb{C}_2^o(k, \geq l, \delta)$.

C. Capacity of 2-Ary k-Runlenth Limited (RLL) Channel

Lemma 1: The capacity of 2-ary k-runlength limited channel is the same as the capacity of 2-ary $(r = k-1)$-constrained channel.

Lemma 2: Let $\boldsymbol{x} \in \Sigma_2^n$. If $\boldsymbol{y} = Diff(\boldsymbol{x})$ is $(r = k-1)$-constrained, then $\boldsymbol{x}$ is 2-ary k-runlength limited [14].

Proof: [Proof of Lemma 1] From Definition 9 and Eq. (10), every binary sequence $\boldsymbol{x}$ and its $\boldsymbol{y} = Diff(\boldsymbol{x})$ have a one-to-one correspondence. And from Lemma 2, if the number of consecutive "0"s of $\boldsymbol{y} = Diff(\boldsymbol{x})$ is no more than $(r = k-1)$, then $\boldsymbol{x}$ is 2-ary k-runlength limited. Furthermore, the number of all binary sequences of length n satisfying 2-ary r-constrained, $N_r(n)$ can be written as follows [16]:

$$N_r(n) = 2^n, 0 < n \leq r, \tag{11}$$

$$N_r(n) = \Sigma_{i=1}^{r+1} N_r(n-i), n > r. \tag{12}$$

According to [14], the number of sequences of length n that satisfy the 2-ary k runlength limited constraint can also be expressed as the recursive relations:

$$|\mathbb{S}_2^{rl}(n,k)| = 2^n, 0 < n \leq k, \tag{13}$$

$$|\mathbb{S}_2^{rl}(n,k)| = \Sigma_{i=1}^{k}(|\mathbb{S}_2^{rl}(n-i,k))|, n > k. \tag{14}$$

Since $r = k-1$, we get that Eqs. (11) and (12) are the same with Eqs. (13) and (14), i.e., the number of binary sequences that satisfy 2-ary k runlength limited constraint is the same as the number of binary sequences which are 2-ary $(r = k-1)$-constrained. □

D. The Capacity of 2-Ary k-RLL and Strong-(l,δ)-Locally-Balanced Constrained Channel

Theorem 1: For $0 \leq k \leq l$, l is an even integer, $\delta > 0$, the capacity of 2-ary k-RLL and strong-(l,δ)-locally-balanced constrained channel, $\mathbb{C}_2^o(k, \geq l, \delta)$, is the capacity of $(r = k-1)$-constrained and $(M = 2\delta+1)$-RDS constrained channel, $\mathbb{C}_2^{cRDS}(r = k-1, M = 2\delta+1)$.

Proof: From Lemma 1, it is known that the 2-ary k-RLL constrained sequence and $(r = k-1)$-constrained sequence is a one-to-one correspondence. And we have known that every $(2\delta+1)$-RDS sequence is strong-(l,δ)-locally-balanced for all $\delta > 0$, $\ell \geq 4$. The set $\mathbb{S}_2^{RDS}(n, 2\delta+1)$ is asymptotically optimal for strong-(l,δ)-locally-balanced constraint. More specifically, as described in Eq. (9), $\mathbb{C}_{RDS}(2\delta+1) = \mathbb{C}_2(\geq \ell, \delta)$, for $\delta > 0$, $\ell \geq 4$ [13]. Thus, We can use $\mathbb{C}_2^{cRDS}(r = k-1, M = 2\delta+1)$ to estimate $\mathbb{C}_2^o(k, \geq l, \delta)$.

For the $(r = k-1)$-constrained and $(M = 2\delta+1)$-RDS constrained system, we can define the matrix $D(\lambda)$ of the system with elements

$$d_{ij} = \lambda^{-i-j+M+2} f(i+j-M-2), 1 \leq i, j \leq M+1. \tag{15}$$

where the size of $D(\lambda)$ is $(M+1) * (M+1)$ and

$$f(p) = \begin{cases} 1, & \text{if } 1 \leq p \leq k \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Then by the Perron-Frobenius theorem, given the parameters k and M, the capacity of the constrained channel is calculated by:

$$\mathbb{C}_2^{cRDS}(k-1, M) = \log_2 \lambda_{max}, \tag{17}$$

where λ_{max} is the greatest real root of $\det[D(\lambda) - I] = 0$.

And the capacity of 2-ary k-RLL and strong-(l, δ)-locally-balanced constrained channel is as follows:

$$\mathbb{C}_2^o(k, \geq l, \delta) = \mathbb{C}_2^{cRDS}(k-1, M) = \log_2 \lambda_{max}. \tag{18}$$

□

Theorem 1 provides a simple way to calculate the channel capacity for binary odd sequences satisfying the 2-ary $k-$RLL and strong-(l, δ)-locally-balanced constraints. A specific example is provided below. Its correctness is verified by a standard method for computing the capacity of a restricted channel, the Perron-Frobenius theorem [16].

E. Case Study

Here, we present a case study to show how the channel capacity is calculated. For $k-1 = 2(k=3)$, $M = 2\delta+1 = 3(\delta=1)$, we have $D(\lambda) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda^{-1} \\ 0 & 0 & \lambda^{-1} & \lambda^{-2} \\ 0 & \lambda^{-1} & \lambda^{-2} & \lambda^{-3} \end{pmatrix}$, and

$$\begin{aligned} \det[D(\lambda) - I] &= \begin{vmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & \lambda^{-1} \\ 0 & 0 & \lambda^{-1}-1 & \lambda^{-2} \\ 0 & \lambda^{-1} & \lambda^{-2} & \lambda^{-3}-1 \end{vmatrix} \\ &= \lambda^2 - \lambda - 1 = 0. \end{aligned}$$

It has two real roots, namely, $\lambda_1 = -0.618$ and $\lambda_2 = 1.618$. Obviously, the $\lambda_{max} = \lambda_2 = 1.618$, and the capacity of the 2-constrained and 3-RDS constrained channel is calculated by:

$$\mathbb{C}_2^{cRDS}(2,3) = \log_2(1.618) = 0.6942,$$

which is also the capacity of 2-ary 3-runlength limited and strong-(l,1)-locally-balanced constrained channel. We can find that this capacity is the same as the capacity of a channel with only the strong-(l,1)-locally-balanced constraint.

For $k-1 = 2(k=3)$, $M = 2\delta+1 = 3(\delta=1)$, this means that, on the one hand, the sequence passing through the channel needs to satisfy the runlength of the same symbol of up to 3; on the other hand, for the strong-(l,δ)-locally-balanced constraint, when $l = 4, \delta = 1$, the sequences need to satisfy the Hamming weight of each window of length $l' \geq 4$ between $[\frac{l'}{2} - 1, \frac{l'}{2} + 1]$. We can see that the sequence only needs to satisfy the strong-(4,1)-locally-balanced constraint.

The constrained graph G that satisfies the strong-(4,1)-locally-balanced constraint is shown in Fig. 2. The circles marked 0, 1, 2, and 3 represent four states, respectively. There are edges between the states, and there are marked output values on the edges. The adjacency matrix of the constrained graph G is: $A_G = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$.

Fig. 2. The constrained graph G for strong-(4,1)-locally-balanced constraint.

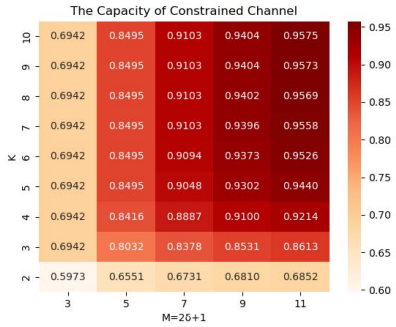

Fig. 3. The capacity of 2-ary k-RLL and strong-(l,δ)-locally-balanced constrained channel for k and $M = 2\delta + 1$.

After calculation, the eigenvalues of the matrix are $\lambda_1 = \frac{\sqrt{5}+1}{2}$, $\lambda_2 = -\frac{\sqrt{5}+1}{2}$, $\lambda_3 = \frac{\sqrt{5}-1}{2}$, $\lambda_4 = -\frac{\sqrt{5}-1}{2}$, and the maximum eigenvalue is $\lambda_{max} = \lambda_1 = \frac{\sqrt{5}+1}{2}$. Then by Perron-Frobenius theorem, we have the capacity is $\mathbb{C}_2(\geq 4, 1) = \log_2(\lambda_{max}) = \log_2(\frac{\sqrt{5}+1}{2}) = 0.6942$. And the value is the same as the capacity of $\mathbb{C}_2^o(3, \geq 4, 1)$. Therefore, with this example, we can verify that the proposed theorem is correct. The results of capacity, $\mathbb{C}_2^o(k, \geq l, \delta)$, of k-RLL and strong-(l,δ)-locally-balanced constrained channel for some parameters that satisfy the conditions are shown in Fig. 3. Then the capacity of the constrained DNA channel can be calculated by $\mathbb{C}_{DNA}(k, \geq l, \delta) = 1 + \mathbb{C}_2^o(k, \geq l, \delta)$.

F. Discussion

In the process of calculating the capacity of 2-ary odd sequences, Eq. (18) indicates that the value of the constrained channel capacity depends only on k and δ and is independent of the length of the sequence n. The choice of parameter l is reduced in the process. In addition, if $M = 2\delta + 1$ remains the same and k is quite large, the k-RLL constraint on the sequence is relatively weak, and the k-RLL constraint has no effect on the strong-(l,δ)-locally-balanced constraint. As a result, the capacity value of 2-ary odd sequences remains the same as the capacity of strong-(l,δ)-locally-balanced constraint, which is independent of l. This is consistent with the following fact:

$$\mathbb{C}_2^o(k, \geq l, \delta) \leq \min\{\mathbb{C}_2(k), \mathbb{C}_{RDS}(2\delta + 1)\}. \quad (19)$$

More importantly, we have provided a calculation method for the capacity of k-RLL and strong-(l,δ)-locally-GC-balanced constrained DNA channel by $\mathbb{C}_{DNA}(k, \geq l, \delta) = 1 + \mathbb{C}_2^o(k, \geq l, \delta)$, combined with **Theorem 1**. This work is the first to obtain the values of the capacity of k-RLL and strong-(l,δ)-locally-GC-balanced constrained DNA channel. And it is of great significance to study the capacity of constrained channels for coding schemes, because the capacity $\mathbb{C}$ is the upper bound on the rate at which any encoding scheme can convert arbitrary data into a sequence with given constraints. If the block encoder is used for binary encoding, then this result can be used as a reference indicator for the encoding rate of this encoder.

IV. Conclusion

We provided a method to calculate the capacity of k-RLL and strong-(l,δ)-locally-GC-balanced constrained channel in DNA storage. In particular, we provided a relationship between the 4-ary constrained channel capacity and the 2-ary constrained channel capacity, which made it simpler to calculate the 4-ary constrained channel capacity. Furthermore, the corresponding binary constrained channel capacity is calculated by theorem, and by the relationship, the capacity of the limited channel based on DNA storage can be obtained. This is the first effort to combine these two constraints. And it is the first to obtain accurate values for the constrained channel capacity. This constrained channel allows DNA data storage system to be less prone to errors during synthesis and sequencing and improves the success rate of PCR amplification. Finally, this result can provide a reliable basis and evaluation index for designing coding schemes in the future.

References

[1] N. Goldman et al., "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, pp. 77–80, Jan. 2013.
[2] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, 2017.
[3] L. C. Meiser et al., "Reading and writing digital data in DNA," *Nature Protocols*, vol. 15, no. 1, pp. 86–101, 2020.
[4] J. Koch, S. Gantenbein, K. Masania, W. J. Stark, Y. Erlich, and R. N. Grass, "A DNA-of-things storage architecture to create materials with embedded memory," *Nature Biotechnol.*, vol. 38, no. 1, pp. 39–43, Jan. 2020.
[5] W. H. Press, J. A. Hawkins, S. K. Jones, J. M. Schaub, and I. J. Finkelstein, "HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 31, pp. 18489–18496, Aug. 2020.
[6] S. Shankland. (Jul. 2, 2019). *Startup Packs All 16Gb of Wikipedia Onto DNA Strands to Demonstrate New Storage Tech*. [Online]. Available: https://www.cnet.com/tech/computing/startup-packs-all-16gb-wikipedia-onto-dna-strands-demonstrate-new-storage-tech/
[7] N. Roquet et al., "DNA-based data storage via combinatorial assembly," *bioRxiv*, Jan. 2021.
[8] M. G. Ross et al., "Characterizing and measuring bias in sequence data," *Genome Biol.*, vol. 14, no. 5, pp. 1–20, 2013.
[9] P. Yakovchuk, "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix," *Nucleic Acids Res.*, vol. 34, no. 2, pp. 564–574, Jan. 2006.
[10] Y. Benita, "Regionalized GC content of template DNA as a predictor of PCR success," *Nucleic Acids Res.*, vol. 31, no. 16, p. e99, Aug. 2003.
[11] K. A. S. Immink and K. Cai, "Properties and constructions of constrained codes for DNA-based data storage," *IEEE Access*, vol. 8, pp. 49523–49531, 2020.
[12] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.
[13] R. Gabrys, H. M. Kiah, A. Vardy, E. Yaakobi, and Y. Zhang, "Locally balanced constraints," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 664–669.
[14] T. T. Nguyen, K. Cai, K. A. S. Immink, and H. M. Kiah, "Capacity-approaching constrained codes with error correction for DNA-based data storage," *IEEE Trans. Inf. Theory*, vol. 67, no. 8, pp. 5602–5613, Aug. 2021.
[15] 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. Oper. Syst.*, Mar. 2016, pp. 637–649.
[16] K. A. S. Immink, *Codes for Mass Data Storage Systems*. Denver, CO, USA: Shannon Foundation, 2004.