On the Capacity of DNA Labeling

Dganit Hanania[1], **Daniella Bar-Lev**[1], **Yevgeni Nogin**[2], **Yoav Shechtman**[2 3 4] and **Eitan Yaakobi**[1]
[1]Department of Computer Science, Technion—Israel Institute of Technology, Haifa 3200003, Israel
[2]Russel Berrie Nanotechnology Institute, Technion, Haifa 320003, Israel
[3]Department of Biomedical Engineering, Technion, Haifa 320003, Israel
[4]Lorry I. Lokey Center for Life Sciences and Engineering, Technion, Haifa 320003, Israel
Email: {dganit ,daniellalev, yaakobi}@cs.technion.ac.il, yevgeni.nogin@gmail.com, yoavsh@bm.technion.ac.il

***Abstract*—*DNA labeling* is a powerful tool in molecular biology and biotechnology that allows for the visualization, detection, and study of DNA at the molecular level. Under this paradigm, a DNA molecule is being *labeled* by specific k patterns and is then imaged. Then, the resulted image is modeled as a $(k+1)$-ary sequence in which any non-zero symbol indicates on the appearance of the corresponding label in the DNA molecule. The primary goal of this work is to study the *labeling capacity*, which is defined as the maximal information rate that can be obtained using this labeling process. The labeling capacity is computed for *any* single label and several results are provided for multiple labels as well. Moreover, we provide the optimal minimal number of labels of length one or two that are needed in order to gain labeling capacity of 2.**

I. Introduction

Labeling of DNA molecules with fluorescent markers is a widely used approach in molecular biology and medicine, with many applications in genomics and microbiology [1]–[3]. This powerful tool allows for the visualization, detection, and study of DNA at the molecular level. Various techniques can be employed to achieve targeted labeling of DNA molecules, such as Fluorescence in situ hybridization (FISH) [1], CRISPR [2], [4] and Methyltransferases [5]. Labeling is also done for other bio-molecules such as proteins and RNA, for applications in sensitive molecular analysis [6], [7] and studying gene expression and regulation [8], respectively.

DNA labeling is used for both specific target sequences and per-base labeling. Per-base labeling is used for DNA sequencing based on sequencing by synthesis (such as illumina sequencing) [9] and Bisulphite sequencing to study of methylation and epigenomics (genetic information beyond the genome sequence) [10]. Target sequence labeling is employed for species identification in clinical microbiology with FISH [11], studying DNA dynamics in living cells [4], optical mapping [12], [13] (for genomic structural variation detection and species identification in microbiology), and the study of DNA-protein interactions, which are fundamental in understanding gene expression and regulation [14]. By attaching a fluorescent label to DNA, researchers can visualize the interaction between DNA and proteins in real-time [14], providing insights into how DNA is packaged, organized, and interacts with proteins in the cell nucleus and how this affects gene expression.

This work takes a first step towards mathematically modeling and analyzing the information rate that can be represented in labeled DNA molecules. More specifically, we study the *labeling capacity*, which refers to the maximum information rate that can be stored by labeling DNA with specific sequence patterns as the labels.

In this work, the labeling process is formally modeled as follows. Assume the DNA sequence is given by $\boldsymbol{x} \in \{A,C,G,T\}^n$ and let $\boldsymbol{\alpha} \in \{A,C,G,T\}^\ell$ be a short sequence which is being used as the *label*. That is, the DNA sequence $\boldsymbol{x}$ is being labeled wherever $\boldsymbol{\alpha}$ appears. As a result, a binary sequence $\boldsymbol{z} \in \{0,1\}^n$ is being received in which $z_i = 1$ if and only if $(x_i,\ldots,x_{i+\ell-1}) = \boldsymbol{\alpha}$. For example, let $\boldsymbol{\alpha} = AC$ be a label of length $\ell = 2$. For $\boldsymbol{x} = AAACGATGACAC$, the received output binary sequence is $\boldsymbol{z} = (0,0,1,0,0,0,0,0,1,0,1,0)$. Clearly, there are other sequences, for example $\boldsymbol{y} = TAACTTTTACAC \neq \boldsymbol{x}$, which result with the same output binary sequence $\boldsymbol{z}$. So, the full capacity is not obtained and the goal of this work is to understand the maximum information rate of this paradigm, and the labeling capacity is referred as the asymptotic ratio between the number of information bits that can be stored and the length n. First, we show that the labeling capacity depends on the length of the used label. For example, for $|\boldsymbol{\alpha}| = 1$, any binary sequence can be achieved by the labeling process and thus the labeling capacity is 1. However, for length-2 labels $\boldsymbol{\alpha} = (\alpha_1,\alpha_2)$, where $\alpha_1 \neq \alpha_2$, binary sequences with two consecutive ones cannot be achieved, i.e., they satisfy the so-called (d,k) *run length limited* (*RLL*) constraint [15] for $(d,k) = (1,\infty)$ and we show that the labeling capacity is the same as the capacity of the $(1,\infty)$ constraint. Besides the label's length, several more properties, such as its periodicity, may affect the labeling capacity. For example, the labeling capacity of AA is larger than the one of the label AC and we extend this result to find the labeling capacity of any label.

The labeling process can also be done using $k > 1$ labels. In this case, the output is a sequence over $\{0,\ldots,k\}$. For example, let $\boldsymbol{\alpha}_1 = AC, \boldsymbol{\alpha}_2 = G$ be two labels. For $\boldsymbol{x} = AAACGATGACAC$ it holds that the received output sequence is $\boldsymbol{z} = (0,0,1,0,2,0,0,2,1,0,1,0)$. The definition of the labeling capacity is extended to multiple labels and we find this capacity when there is no overlap between the labels or for two non-cyclic labels in the special case where there is a unique way for the two labels to overlap each other.

The last part of this work is dedicated to finding the minimal number of needed labels of a given length in order to obtain the maximum labeling capacity 2. For example, for labels of length 1, three different labels are necessary and sufficient to decode every sequence over $\{A,C,G,T\}^n$ and to have

The research was Funded by the European Union (ERC, DNAStorage, 865630). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

978-1-6654-7554-9/23/$31.00 ©2023 IEEE

capacity 2. For labels of length 2, it is clear that achieving labeling capacity 2 may be obtained using any 15 different labels of length two. Our main result here claims that this can be accomplished with 10 labels and no less than 10.

The rest of this paper is organized as follows. Section II formally defines the labeling channel and several useful definitions. Section III calculates the labeling capacity for a single label while considering its periodicity and overlap. In Section IV we extend the results for multiple labels. Lastly, in Section V, the minimum number of labels needed to obtain the full capacity is studied for labels of length one or two. All missing proofs will be presented in the extended version of the paper.

II. Definitions and Preliminaries

Let Σ_q denote the q-ary alphabet $\{0,1,\ldots,q-1\}$. For $q=4$, we mostly refer to the DNA alphabet, that is, $\Sigma_4=\{A,C,G,T\}$. Denote by $[n]$ the set $\{1,2,\ldots,n\}$. For a sequence $\boldsymbol{x}=(x_1,\ldots,x_n)\in\Sigma_q^n$, and $1\le i\le n-k+1$, let $\boldsymbol{x}_{[i;k]}=(x_i,\ldots,x_{i+k-1})$. A *label* $\boldsymbol{\alpha}\in\Sigma_q^\ell$ is a relatively short (typically at most 6 bases long) sequence over Σ_q. Next, the *labeling model* studied in this work is formally defined.

Definition 1. Let $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_k$ be k labels of lengths $\ell_1,\ldots,\ell_k$, respectively. Among the k labels, there is no label that is a prefix of another label. Denote $\underline{\boldsymbol{\alpha}}=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_k)$.

- The ***$\underline{\boldsymbol{\alpha}}$-labeling sequence*** of $\boldsymbol{x}=(x_1,\ldots,x_n)\in\Sigma_q^n$ is the sequence $L_{\underline{\boldsymbol{\alpha}}}(\boldsymbol{x})=(c_1,\ldots,c_n)\in\Sigma_{k+1}^n$, in which $c_i=j$ if $\boldsymbol{x}_{[i;\ell_j]}=\boldsymbol{\alpha}_j$ and $i\le n-\ell_j+1$, and if such j does not exist then $c_i=0$.
- A sequence $\boldsymbol{u}\in\{0,\ldots,k\}^n$ is called a ***valid $\underline{\boldsymbol{\alpha}}$-labeling sequence*** if there exists a sequence $\boldsymbol{x}\in\Sigma_q^n$ such that $\boldsymbol{u}=L_{\underline{\boldsymbol{\alpha}}}(\boldsymbol{x})$.
- Denote by $F_n(\underline{\boldsymbol{\alpha}})$ the set of all valid $\underline{\boldsymbol{\alpha}}$-labeling sequences of length n, which means, the image of the mapping $L_{\underline{\boldsymbol{\alpha}}}$. That is, $F_n(\underline{\boldsymbol{\alpha}})=\{L_{\underline{\boldsymbol{\alpha}}}(\boldsymbol{x})|\boldsymbol{x}\in\Sigma_q^n\}$. Denote the ***labeling capacity*** of $\underline{\boldsymbol{\alpha}}$ by

$$\mathsf{cap}(\underline{\boldsymbol{\alpha}})\triangleq\limsup_{n\to\infty}\frac{\log_2(|F_n(\underline{\boldsymbol{\alpha}})|)}{n}.$$

In case there is only one label that is being used, another equivalent way to represent the labeling sequence is given in the next definition. This equivalent definition is used merely to ease the notations in some of the proofs in the paper.

Definition 2. Let $\boldsymbol{\alpha}$ be a label of length ℓ.

- The ***complete-$\boldsymbol{\alpha}$-labeling sequence*** of $\boldsymbol{x}=(x_1,\ldots,x_n)\in\Sigma_q^n$ is the binary sequence $\widehat{L}_{\boldsymbol{\alpha}}(\boldsymbol{x})=(c_1,\ldots,c_n)$, in which $\boldsymbol{c}_{[i;\ell]}=(1,\ldots,1)$ if $\boldsymbol{x}_{[i;\ell]}=\boldsymbol{\alpha}$ and $i\le n-\ell+1$.
- A sequence $\boldsymbol{u}\in\Sigma_2^n$ is called a ***valid complete-$\boldsymbol{\alpha}$-labeling sequence*** if there exists a sequence $\boldsymbol{x}\in\Sigma_q^n$ such that $\boldsymbol{u}=\widehat{L}_{\boldsymbol{\alpha}}(\boldsymbol{x})$.
- Denote by $\widehat{F}_n(\boldsymbol{\alpha})$ the set of all valid complete-$\boldsymbol{\alpha}$-labeling sequences of length n.

Note that there is a bijection between $F_n(\boldsymbol{\alpha})$ and $\widehat{F}_n(\boldsymbol{\alpha})$. So, in order to compute the labeling capacity, in some of the cases, $\widehat{F}_n(\boldsymbol{\alpha})$ will be computed instead of $F_n(\boldsymbol{\alpha})$, for convenience.

Example 1. For $\boldsymbol{\alpha}_1=CG,\boldsymbol{\alpha}_2=A,\boldsymbol{x}=ACCGGCGATA$, it holds that $L_{(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2)}(\boldsymbol{x})=(2,0,1,0,0,1,0,2,0,2)$. Moreover, $\widehat{L}_{\boldsymbol{\alpha}_1}(\boldsymbol{x})=(0,0,1,1,0,1,1,0,0,0)$ and $L_{\boldsymbol{\alpha}_1}(\boldsymbol{x})=(0,0,1,0,0,1,0,0,0,0)$.

The following definitions will be helpful in order to discuss different types of labels.

Definition 3. Let $\boldsymbol{\alpha},\boldsymbol{\alpha}'$ be a label of length ℓ,ℓ', respectively.

- The ***period*** of $\boldsymbol{\alpha}$ is $\mathcal{P}(\boldsymbol{\alpha})\triangleq\min\{p\in[\ell]:p|\ell,\boldsymbol{\alpha}_{[1;p]}=\boldsymbol{\alpha}_{[(t-1)p+1;p]}\text{ for }t\in[\frac{\ell}{p}]\}$. In case $p=\ell$, there is no period in $\boldsymbol{\alpha}$ and the label is called a ***non-periodic*** label.
- The ***overlap*** between $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'\neq\boldsymbol{\alpha}$ is $\mathcal{O}(\boldsymbol{\alpha},\boldsymbol{\alpha}')\triangleq\max\{r\in[\min\{\ell,\ell'\}]:\boldsymbol{\alpha}'_{[1;r]}=\boldsymbol{\alpha}_{[\ell'-r+1;r]}\}$. In other words, $\mathcal{O}(\boldsymbol{\alpha},\boldsymbol{\alpha}')$ is the maximal size of a suffix of $\boldsymbol{\alpha}$ which is identical to a prefix of $\boldsymbol{\alpha}'$. In case $\mathcal{O}(\boldsymbol{\alpha},\boldsymbol{\alpha}')$ does not exist, we define $\mathcal{O}(\boldsymbol{\alpha},\boldsymbol{\alpha}')\triangleq 0$. The labels $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ are called ***overlapping labels*** if $\mathcal{O}(\boldsymbol{\alpha},\boldsymbol{\alpha}')>0$ or $\mathcal{O}(\boldsymbol{\alpha}',\boldsymbol{\alpha})>0$.
- The ***cyclic overlap*** of $\boldsymbol{\alpha}$ is $\mathcal{O}(\boldsymbol{\alpha})\triangleq\mathcal{O}(\boldsymbol{\alpha}_{[1;\ell-1]},\boldsymbol{\alpha}_{[2;\ell-1]})$ if $\ell>1$, and otherwise $\mathcal{O}(\boldsymbol{\alpha})\triangleq 0$. In case $\mathcal{O}(\boldsymbol{\alpha})=0$, $\boldsymbol{\alpha}$ is called a ***non-cyclic label***.

Note that a periodic label is also a cyclic label but a cyclic label is not necessarily periodic. For a periodic label $\boldsymbol{\alpha}$ of length ℓ, it holds that $\mathcal{O}(\boldsymbol{\alpha})=\ell-\mathcal{P}(\boldsymbol{\alpha})$. The next example exemplifies the definitions above.

Example 2. The labels $\boldsymbol{\alpha}_1=CGCGCG$ and $\boldsymbol{\alpha}_2=GATG$ are overlapping labels. It holds that $\mathcal{O}(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2)=1$ and $\mathcal{O}(\boldsymbol{\alpha}_2,\boldsymbol{\alpha}_1)=0$. Moreover, $\mathcal{O}(\boldsymbol{\alpha}_1)=4$ and $\mathcal{P}(\boldsymbol{\alpha}_1)=2$. In contrast, $\mathcal{O}(\boldsymbol{\alpha}_2)=1$ but it is a non-periodic label.

One of the goals in this work is to calculate the labeling capacity using one or more labels. Some of the results will be derived by drawing a connection to constrained systems. In order to establish this connection, several more definitions are introduced as described in [15].

Definition 4. A ***finite labeled***[1] ***directed graph*** $G=(V,E,L)$ is a graph which consists of a finite set of states V, a finite set of edges E, and an edge labeling $L:E\to\Sigma_q$. A sequence $\boldsymbol{w}$ over Σ_q is ***generated*** by π (and G) if π is a path in G which is labeled by the sequence $\boldsymbol{w}$. A labeled graph G is ***deterministic*** if the outgoing edges from each state are labeled distinctly. A ***constraint*** S is the set of all sequences over Σ_q that are generated by a labeled graph G. In this case, it is said that G ***presents*** S and it is denoted by $S=S(G)$. Denote the set of sequences of length n in the constraint S by $S(n)=|S\cap\Sigma_q^n|$. It is known that for each constraint, there exists a deterministic graph that presents it. The ***capacity*** of a constraint S is

$$\mathsf{cap}(S)\triangleq\limsup_{n\to\infty}\frac{1}{n}\log_2(S(n)).$$

For a deterministic presentation G of S, it holds that $\mathsf{cap}(S)=\log_2(\lambda(A_G))$, where $\lambda(A_G)$ is the spectral radius (Perron eigenvalue) which is the largest real eigenvalue in absolute value of the adjacency matrix of G.

Some of the results in the paper will be connected to a specific constraint, known as the the run-length limited (RLL) constraint, as described in the next definition.

Definition 5. A sequence over Σ_q satisfies the ***(q,d,k)-RLL constraint*** if between every two consecutive non-zero symbols there are at least d zeroes and there is no run of zeroes of length $k+1$. Denote the set of all sequences of length n that

[1]Contrary to the definition of labeling in this work, here the meaning of labeling is giving labels to the edges of the graph.

satisfy the (q,d,k)-RLL constraint by $\mathcal{C}_{q,d,k}(n)$. For the case of $q=2$, this constraint is called the (d,k)-***RLL constraint***, and $\mathcal{C}_{2,d,k}$ will be denoted by $\mathcal{C}_{d,k}$.

It has been proven that $\mathsf{cap}(\mathcal{C}_{M,d,\infty}) = \log_2\lambda$ when λ is the largest real root of $x^{d+1}-x^d-(M-1)$ [16]. Lastly, for $S\subseteq\Sigma_q^{n_1}$, $\boldsymbol{u}\in\Sigma_q^{n_2}$, let $S\circ\boldsymbol{u}\triangleq\{\boldsymbol{w}\in\Sigma_q^{n_1+n_2}|\exists\boldsymbol{s}\in S,\boldsymbol{w}=\boldsymbol{s}\boldsymbol{u}\}$.

In this work, the following problems will be solved.

Problem 1. The k-labeling capacity problem: Let $\underline{\boldsymbol{\alpha}}=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_k)$ be k labels that are being used in order to label sequences over Σ_q^n. Find the labeling capacity $\mathsf{cap}(\underline{\boldsymbol{\alpha}})$.

Problem 2. Let $\underline{\boldsymbol{\alpha}}=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_s)$ be s labels of length $\ell\geq 1$. Find the minimal s such that "almost" every $\boldsymbol{x}\in\Sigma_q^n$ can be determined given its $\underline{\boldsymbol{\alpha}}$-labeling sequence. More specifically, we are interested in computing the minimal number of labels of a fixed length ℓ that are needed in order to gain capacity of two. Mathematically, the problem is to find the value of

$$s(\ell)\triangleq\min\{s\in\mathbb{N}|\exists\underline{\boldsymbol{\alpha}}=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_s)\in\Sigma_{q^\ell}^s,\mathsf{cap}(\underline{\boldsymbol{\alpha}})=2\}.$$

III. The Labeling Capacity of a Single Label

In this section, we provide a full solution to the labeling capacity in case a single label is used. The next theorem solves the case where the label is non-cyclic.

Theorem 1. Let $\boldsymbol{\alpha}$ be a non-cyclic label of length ℓ. Then, $\mathsf{cap}(\boldsymbol{\alpha})=\mathsf{cap}(\mathcal{C}_{\ell-1,\infty})$. That is, $\mathsf{cap}(\boldsymbol{\alpha})=\log_2\lambda$ when λ is the largest real root of $x^\ell-x^{\ell-1}-1$.

Proof: Let $\boldsymbol{\alpha}$ be a non-cyclic label of length ℓ. For $\boldsymbol{x}\in\Sigma_q^n$, let $\boldsymbol{y}\in\Sigma_2^n$ be the $\boldsymbol{\alpha}$-labeling sequence of $\boldsymbol{x}$, i.e., $\boldsymbol{y}\triangleq L_{\boldsymbol{\alpha}}(\boldsymbol{x})$. By definition, it holds that $y_i=1$ if and only if $\boldsymbol{x}_{[i;\ell]}=\boldsymbol{\alpha}$. Since the label $\boldsymbol{\alpha}$ is non-cyclic it holds that if $\boldsymbol{x}_{[i;\ell]}=\boldsymbol{\alpha}$, then $\boldsymbol{x}_{[j;\ell]}\neq\boldsymbol{\alpha}$ for $i+1\leq j\leq i+\ell-1$. Hence, for every $\boldsymbol{\alpha}$-labeling sequence it holds that after each one there are at least $\ell-1$ zeroes, i.e., $\boldsymbol{y}$ satisfies the $(\ell-1,\infty)$-RLL constraint and ends with $\ell-1$ zeros. So, $F_n(\boldsymbol{\alpha})\subseteq\mathcal{C}_{\ell-1,\infty}(n-(\ell-1))\circ 0^{\ell-1}$. In order to prove inclusion in the other direction, let $\boldsymbol{u}\in\mathcal{C}_{\ell-1,\infty}(n-\ell+1)\circ 0^{\ell-1}$. Let $\boldsymbol{v}\in\Sigma_q^n$ such that $\boldsymbol{v}_{[i;\ell]}=\boldsymbol{\alpha}$ if and only if $u_i=1$. It holds that $L_{\boldsymbol{\alpha}}(\boldsymbol{v})=\boldsymbol{u}$. From the definition of $\mathcal{C}_{\ell-1,\infty}(n-(\ell-1))\circ 0^{\ell-1}$, after each one in $\boldsymbol{u}$ there are at least $\ell-1$ zeros and $|\boldsymbol{\alpha}|=\ell$, so such a $\boldsymbol{v}$ exists. Hence, for $n\geq\ell-1$, $|F_n(\boldsymbol{\alpha})|=|\mathcal{C}_{\ell-1,\infty}(n-\ell+1)\circ 0^{\ell-1}|$. So, $\mathsf{cap}(\boldsymbol{\alpha})=\limsup_{n\to\infty}\frac{\log_2(|F_n(\boldsymbol{\alpha})|)}{n}=\limsup_{n\to\infty}\frac{\log_2(|\mathcal{C}_{\ell-1,\infty}(n-\ell+1)|)\cdot(n-\ell+1)}{n\cdot(n-\ell+1)}=\limsup_{n\to\infty}\frac{\mathsf{cap}(\mathcal{C}_{\ell-1,\infty})\cdot(n-\ell+1)}{n}=\mathsf{cap}(\mathcal{C}_{\ell-1,\infty})=\log_2\lambda$, where λ is the largest real root of $x^\ell-x^{\ell-1}-1$. ∎

Before we continue with the labeling capacity for periodic labels, the next example motivates the solution of this case.

Example 3. Let $\boldsymbol{\alpha}=CGCG$, so $\mathcal{P}(\boldsymbol{\alpha})=2$. Let $\boldsymbol{x}\in\{A,C,G,T\}^n$ and let $\boldsymbol{y}\in\Sigma_2^n$ be the complete-$\boldsymbol{\alpha}$-labeling sequence of $\boldsymbol{x}$, i.e., $\boldsymbol{y}\triangleq\widehat{L}_{(\boldsymbol{\alpha})}(\boldsymbol{x})$. It holds that if $\boldsymbol{x}_{[i;4]}=CGCG$, then $y_{[i,4]}=(1,1,1,1)$. So, if for $k\geq 2$, $\boldsymbol{x}_{[i;2k]}$ consists of a run of k CGs and $\boldsymbol{x}_{[i+2k;2]}\neq CG$, then $\boldsymbol{y}_{[i;2k]}=(1,1,\ldots,1)$ and $y_{i+2k}=0$, because otherwise this implies that $\boldsymbol{x}_{[i+2k;4]}=CGCG$ but $\boldsymbol{x}_{[i+2k;2]}\neq CG$. It can be concluded that every valid complete-$\boldsymbol{\alpha}$-labeling sequence is a binary sequence in which the length of every run of ones is even and at least four. Denote this set by $S_{E\geq 4}$ and $S_{E\geq 4}(n)\triangleq S_{E\geq 4}\cap\Sigma_2^n$. After proving that $\widehat{F}_n(\boldsymbol{\alpha})\subseteq S_{E\geq 4}(n)$, in order to prove equality between those sets, we prove inclusion in the other direction next. Let $\boldsymbol{u}\in S_{E\geq 4}(n)$ and let $\boldsymbol{v}\in\Sigma_4^n$ be such that for $k\geq 2$, $\boldsymbol{v}_{[i;2k]}=CGCG\cdots CG$ if $\boldsymbol{u}_{[i;2k]}=(1,\ldots,1)$ and $u_{i-1}=0$ or $i=0$. It holds that $\widehat{L}_{\boldsymbol{\alpha}}(\boldsymbol{v})=\boldsymbol{u}$.

Hence, we have that $|\widehat{F}_n(\boldsymbol{\alpha})|=|S_{E\geq 4}(n)|$. Denote the constraint that is presented in the graph in Figure 1 by S.

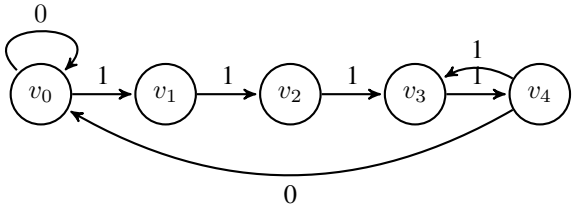


Fig. 1: The Constrained Graph for Example 3.

From the structure of the graph, for any sequence $\boldsymbol{y}'$ of length $n-6$, when $n\geq 6$, that is generated by the graph, there exist sequences $\boldsymbol{y}'',\boldsymbol{y}'''\in\Sigma_2^3$, such that $\boldsymbol{y}\triangleq\boldsymbol{y}'''\circ\boldsymbol{y}'\circ\boldsymbol{y}''\in S_{E\geq 4}(n)\subseteq S(n)$. Hence, we have an injection from $S(n-6)$ to $S_{E\geq 4}(n)$ and $|S(n-6)|\leq|S_{E\geq 4}(n)|\leq|S(n)|$. Hence, $\mathsf{cap}(\underline{\boldsymbol{\alpha}})=\mathsf{cap}(S)$. The adjacency matrix of this graph is

$$\begin{pmatrix}1&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\\1&0&0&1&0\end{pmatrix},$$

and its characteristic polynomial is $x^5-x^4-x^3+x^2-1$. Thus, the capacity of S is $\log_2(\lambda)$ when $\lambda\approx 1.44$ is the largest real root of this polynomial.

This example will lead to the general case of calculating the labeling capacity of a label with period $p<\ell$.

Theorem 2. Let $\boldsymbol{\alpha}$ be a label of length ℓ with $\mathcal{P}(\boldsymbol{\alpha})<\ell$. It holds that $\mathsf{cap}(\boldsymbol{\alpha})=\log_2(\lambda)$ when λ is the largest real root of the polynomial $x^{\ell+1}-x^\ell-x^{\ell-p+1}+x^{\ell-p}-1$.

Proof: Let $\boldsymbol{\alpha}$ be a label of length ℓ with $p=\mathcal{P}(\boldsymbol{\alpha})<\ell$. Denote $\boldsymbol{\alpha}'=\boldsymbol{\alpha}_{[1;p]}$. Let $\boldsymbol{x}\in\Sigma_q^n$ and let $\boldsymbol{y}\triangleq\widehat{L}_{\boldsymbol{\alpha}}(\boldsymbol{x})$. It holds that if $\boldsymbol{x}_{[i;\ell]}=\boldsymbol{\alpha}$, then $\boldsymbol{y}_{[i;\ell]}=(1,\ldots,1)$. So, if there exists $k\geq\frac{\ell}{p}$ such that $\boldsymbol{x}_{[i;pk]}$ consists of the concatenation of $\boldsymbol{\alpha}'$ k times and $\boldsymbol{x}_{[i+pk;p]}\neq\boldsymbol{\alpha}'$, then $\boldsymbol{y}_{[i;pk]}=(1,\ldots,1)$ and $y_{i+pk}=0$, since otherwise this implies that $\boldsymbol{x}_{[i+pk;\ell]}=\boldsymbol{\alpha}$ but $\boldsymbol{x}_{[i+pk;p]}\neq\boldsymbol{\alpha}'$. It can be concluded that every valid complete-$\boldsymbol{\alpha}$-labeling sequence is a binary sequence in which the length of every run of ones is at least ℓ and divisible by p. Denote this set by $S_{p\geq\ell}$ and $S_{p\geq\ell}(n)\triangleq S_{p\geq\ell}\cap\Sigma_2^n$. After proving that $\widehat{F}_n(\boldsymbol{\alpha})\subseteq S_{p\geq\ell}(n)$, in order to prove equality between those sets, we prove inclusion in the other direction next. Let $\boldsymbol{u}\in S_{p\geq\ell}(n)$ and let $\boldsymbol{v}\in\Sigma_q^n$ be a sequence in which if for $k\geq\frac{\ell}{p}$, $\boldsymbol{u}_{[i,pk]}=(1,\ldots,1)$ and $u_{i-1}=0$ or $i=0$, then $\boldsymbol{v}_{[i,pk]}=\boldsymbol{\alpha}'\cdots\boldsymbol{\alpha}'$. It holds that $\widehat{L}_{\boldsymbol{\alpha}}(\boldsymbol{v})=\boldsymbol{u}$. Hence, it holds that $|\widehat{F}_n(\boldsymbol{\alpha})|=|S_{p\geq\ell}(n)|$. Denote the constraint that is presented in Figure 2 by S, where the missing edges of the graph are labeled with 1.

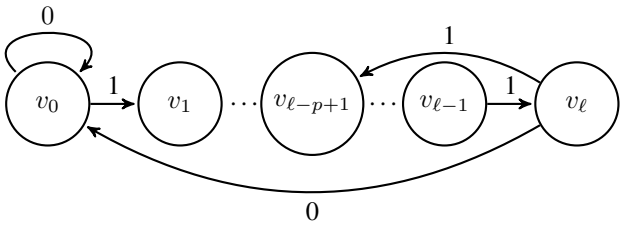


Fig. 2: Graph for Theorem 2.

From the structure of the graph, for any sequence $\boldsymbol{y}'$ of length $n-2(\ell-1)$, when $n \geq 2(\ell-1)$, that is generated by the graph, there exist sequences $\boldsymbol{y}'', \boldsymbol{y}''' \in \Sigma_2^{\ell-1}$, such that $\boldsymbol{y} \triangleq \boldsymbol{y}''' \circ \boldsymbol{y}' \circ \boldsymbol{y}'' \in S_{p\geq\ell}(n) \subseteq S(n)$. Hence, we have an injection from $S(n-2(\ell-1))$ to $S_{p\geq\ell}(n)$ and $|S(n-2(\ell-1))| \leq |S_{p\geq\ell}(n)| \leq |S(n)|$. It implies that $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = \mathsf{cap}(S)$. The adjacency matrix of this graph is

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & 1 & \cdots & 0 \end{pmatrix},$$

where $A_{i,j} = 1$ for (i,j) such that: $i = 0, j = 0$ or $i = j-1$ or $i = \ell, j = 0$ or $i = \ell, j = \ell-p+1$ and otherwise, $A_{i,j} = 0$. It can be shown that the characteristic polynomial of this matrix is $x^{\ell+1} - x^\ell + x^{\ell-p} - x^{\ell-p+1} - 1$. Thus, $\mathsf{cap}(S) = \log_2(\lambda)$ where λ is the largest real root of the last polynomial. ∎

Lastly, we examine the case of a non-periodic label with a cyclic overlap $r > 0$. An example and the proof for the general case of a label with a non-periodic overlap will be presented in the extended version of the paper. The case of a label with a periodic cyclic overlap can be studied in a similar way.

Theorem 3. Let $\boldsymbol{\alpha}$ be a non-periodic label of length ℓ with a non-periodic cyclic overlap $r > 0$. It holds that $\mathsf{cap}(\boldsymbol{\alpha}) = \log_2(\lambda)$ when λ is the largest real root of the polynomial $x^\ell - x^{\ell-1} - x^r + x^{r-1} - 1$.

IV. The Labeling Capacity of Multiple Labels

In this section we study the labeling capacity in case multiple labels are being used. The next theorem solves the case of k non-overlapping non-cyclic labels.

Theorem 4. Let $\underline{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k)$ be k non-overlapping non-cyclic labels of lengths $\ell_1 \leq \cdots \leq \ell_k$ respectively. Denote the number of labels of length j by m_j. It holds that $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = \log_2(\lambda)$ where λ is the largest real root of the polynomial $x^{\ell_k} - (1+m_1)x^{\ell_k-1} - \sum_{i=2}^{\ell_k} m_i x^{\ell_k-i}$.

Proof: The valid $\underline{\boldsymbol{\alpha}}$-labeling sequences are the sequences over Σ_{k+1} in which after each $i \in [k]$ there are at least $\ell_i - 1$ zeroes. The valid $\underline{\boldsymbol{\alpha}}$-labeling sequences could be presented in a graph G, for which the adjacency matrix $A_{\ell_k \times \ell_k}$ will be as follows. For (i,j) such that $i = j+1$, $A_{i,j} = 1$. Moreover, $A_{0,0} = 1 + m_1$ and for $j > 0$, $A_{0,j} = m_{j+1}$. Otherwise, $A_{i,j} = 0$. It can be shown that the characteristic polynomial of this matrix is $x^{\ell_k} - (1+m_1)x^{\ell_k-1} - \sum_{i=2}^{\ell_k} m_i x^{\ell_k-i}$. So, it holds that $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = \mathsf{cap}(S(G)) = \log_2(\lambda)$, where λ is the largest real root of this polynomial. ∎

Note that in case the k non-overlapping non-cyclic labels are of the same length ℓ, $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = \log_2 \lambda$ when λ is the largest real root of $x^\ell - x^{\ell-1} - k$. It has been proven in [16] that this is the capacity of the $(k+1, \ell-1, \infty)$-RLL constraint.

The last case to be discussed is the case of using overlapping labels. In this paper, the labeling capacity of two non-cyclic labels $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$, when $\mathcal{O}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) > 0, \mathcal{O}(\boldsymbol{\alpha}_2, \boldsymbol{\alpha}_1) = 0$ will be provided. Additional cases can be studied in a similar way as is done in the following example and theorem.

Example 4. Let $\underline{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ when $\boldsymbol{\alpha}_1 = ACGT, \boldsymbol{\alpha}_2 = GTT$, are two non-cyclic labels of lengths $\ell_1 = 4, \ell_2 = 3$. It holds that $\mathcal{O}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = t = 2, \mathcal{O}(\boldsymbol{\alpha}_2, \boldsymbol{\alpha}_1) = 0$. Let $\boldsymbol{x} \in \Sigma_4^n$ and let $\boldsymbol{y} \in \Sigma_3^n$ be the $\underline{\boldsymbol{\alpha}}$-labeling sequence of $\boldsymbol{x}$, i.e., $\boldsymbol{y} \triangleq L_{\underline{\boldsymbol{\alpha}}}(\boldsymbol{x})$. From the definition, it holds that:

- If $\boldsymbol{x}_{[i;4]} = ACGT$ and $x_{i+4} \neq T$, then $\boldsymbol{y}_{[i;4]} = (1,0,0,0)$. In the general case, if $\boldsymbol{x}_{[i;\ell_1]} = \boldsymbol{\alpha}_1$ and $\boldsymbol{x}_{[i+\ell_1;\ell_2-t]} \neq \boldsymbol{\alpha}_{2[1+t;\ell_2-t]}$, then $\boldsymbol{y}_{[i;\ell_1]} = (1,0,\ldots,0)$.
- If $\boldsymbol{x}_{[i;4]} = ACGT$ and $x_{i+4} = T$, then $\boldsymbol{y}_{[i;5]} = (1,0,2,0,0)$. In general, if $\boldsymbol{x}_{[i;\ell_1]} = \boldsymbol{\alpha}_1$ and $\boldsymbol{x}_{[i+\ell_1;\ell_2-t]} = \boldsymbol{\alpha}_{2[1+t;\ell_2-t]}$, then $\boldsymbol{y}_{[i;\ell_1-t]} = (1,0,\ldots,0)$, $\boldsymbol{y}_{[i+\ell_1-t;\ell_2]} = (2,0,\ldots,0)$.
- If $\boldsymbol{x}_{[i;3]} = GTT$ then $\boldsymbol{y}_{[i;3]} = (2,0,0)$. In general, if $\boldsymbol{x}_{[i;\ell_2]} = \boldsymbol{\alpha}_2$ then $\boldsymbol{y}_{[i;\ell_2]} = (2,0,\ldots,0)$.
- Else, $y_i = 0$.

The correctness is due to the fact that the sequences are non-cyclic. So, every valid $\underline{\boldsymbol{\alpha}}$-labeling sequence is a ternary sequence in which (1) each one is followed by three zeroes or zero and two, and (2) each two is followed by at least two zeroes.

Denote the set of ternary sequences that hold these two conditions by S_c and $S_c(n) \triangleq S_c \cap \Sigma_3^n$. After proving that $F_n(\underline{\boldsymbol{\alpha}}) \subseteq S_c(n)$, in order to prove equality between those sets, we prove inclusion in the other direction next. Let $\boldsymbol{u} \in S_c(n)$ and let $\boldsymbol{v} \in \Sigma_4^n$ be a sequence in which if $\boldsymbol{u}_i = 1$, then $\boldsymbol{v}_{[i;4]} = ACGT$ and if $\boldsymbol{u}_i = 2$, then $\boldsymbol{v}_{[i;3]} = GTT$. It holds that $L_{\underline{\boldsymbol{\alpha}}}(\boldsymbol{v}) = \boldsymbol{u}$. So, the valid $\underline{\boldsymbol{\alpha}}$-labeling sequences of length n are the sequences in $S_c(n)$, which means that $F_n(\underline{\boldsymbol{\alpha}}) = S_c(n)$. Denote the constraint that is presented in Figure 3 by S.

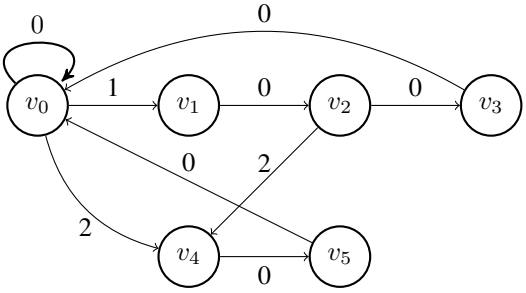


Fig. 3: Graph for Example 4.

From the structure of the graph, for any sequence $\boldsymbol{y}'$ of length $n-3$, when $n \geq 3$, that is generated by the graph, it holds that $\boldsymbol{y} \triangleq \boldsymbol{y}' \circ 000 \in F_n(\underline{\boldsymbol{\alpha}}) \subseteq S(n)$. So, $|S(n-3)| \leq |F_n(\underline{\boldsymbol{\alpha}})| \leq |S(n)|$. As a result, $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = \mathsf{cap}(S)$. The adjacency matrix of this graph is

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

which has the characteristic polynomial $x^6-x^5-x^3-x^2-x$. Thus, the capacity of S is $\log_2(\lambda)$ when $\lambda \approx 1.685$ is the largest real root of this polynomial.

Theorem 5. Let $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ be two non-cyclic labels of lengths ℓ_1, ℓ_2 respectively when $\mathcal{O}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = t > 0, \mathcal{O}(\boldsymbol{\alpha}_2, \boldsymbol{\alpha}_1) = 0$. It holds that $\mathsf{cap}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \log_2(\lambda)$, where λ is the largest real root of $x^{\ell_1+\ell_2-1} - x^{\ell_1+\ell_2-2} - x^{\ell_1-1} - x^{\ell_2-1} - x^{t-1}$.

The proof for this theorem is similar to the previous example and will be presented in the extended version of the paper.

V. The Minimal Number of Labels Problem

In this section, we solve Problem 2 for $\ell = 1, 2$. That is, we find the minimal number of labels of length ℓ that are needed in order to gain labeling capacity of two.

Theorem 6. It holds that $s(1) = 3$.

Proof: Let $\underline{\boldsymbol{\alpha}} = (A, C, G)$. The valid $\underline{\boldsymbol{\alpha}}$-labeling sequences are all the sequences over Σ_4^n. So, $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = \log_2(4) = 2$. Additionally, w.l.o.g, let $\underline{\boldsymbol{\alpha}} = (A, C)$. The valid $\underline{\boldsymbol{\alpha}}$-labeling sequences are all the sequences over Σ_3^n. So, $\mathsf{cap}(\boldsymbol{\alpha}) = \log_2(3) < 2$. ∎

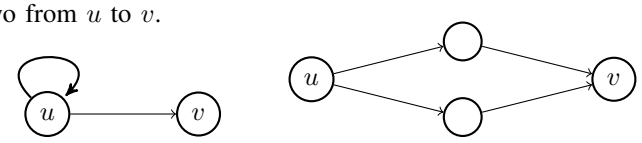
Our main result in this section is stated next.

Theorem 7. It holds that $s(2) = 10$.

In order to prove this theorem, first, consider the following definition and theorem.

Definition 6. Let $G = (V, E)$ be a directed graph. This graph is said to be a ***path unique graph*** if for any $k \geq 1$, between any two vertices there exists at most one path of length k.

Example 5. The graph in Figure 4a is a path unique graph since there are no paths from v to u and every different path from u to v is of different length. The graph in Figure 4b is not a path unique graph since there are two paths of length two from u to v.



(a) (b)

Fig. 4: Examples for a path unique graph (a) and a graph which is not path unique (b).

Theorem 8. Let $\mathcal{S}$ be a subset of the labels of length two. Let $G = (V, E)$ be a directed graph in which $V = \Sigma_4$ and $E = \{(x, y) | xy \in \mathcal{S}\}$. Denote $\bar{\mathcal{S}} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{16-|\mathcal{S}|}\}$ the set of labels of length two which are not in $\mathcal{S}$ and $\underline{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{16-|\mathcal{S}|})$. It holds that $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = 2$ if and only if G is a path unique graph.

Proof: Let $\mathcal{S}$ be a subset of labels of length two, let $\bar{\mathcal{S}} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{16-|\mathcal{S}|}\}$ be the set of labels of length two which are not in $\mathcal{S}$ and $\underline{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{16-|\mathcal{S}|})$. Let $G = (V, E)$ be a directed graph in which $V = \Sigma_4$ and $E = \{(x, y) | xy \in \mathcal{S}\}$. Additionally, for $\boldsymbol{x} \in \Sigma_4^n$ let $\boldsymbol{y} \in \Sigma_{1+|\bar{\mathcal{S}}|}^n$ be the $\underline{\boldsymbol{\alpha}}$-labeling sequence of $\boldsymbol{x}$, so, $y_i = i$ if and only if $\boldsymbol{x}_{[i;2]} = \boldsymbol{\alpha}_i$.

First, assume that G is a path unique graph. Note that from the definition of path unique graph, G is not a full graph, which means that $\bar{\mathcal{S}}$ is not an empty set. Let $ab \in \bar{\mathcal{S}}$. In order to prove that $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = 2$, it will be shown that the capacity of the sequences that start and end with ab is 2, and so $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = 2$. It will be shown that using sequences of length n that start and end with ab, the mapping $L_{\underline{\boldsymbol{\alpha}}}$ is one-to-one, which means that the number of valid $\underline{\boldsymbol{\alpha}}$-labeling sequences of length n is 4^{n-4} and the labeling capacity is $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = \limsup_{n\to\infty} \frac{\log_2(4^{n-4})}{n} = 2$.

Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Sigma_4^n$ be sequences that start and end with ab such that $\boldsymbol{y} = L_{\underline{\boldsymbol{\alpha}}}(\boldsymbol{x}_1) = L_{\underline{\boldsymbol{\alpha}}}(\boldsymbol{x}_2) \neq (0, \ldots, 0)$. It will be shown that $\boldsymbol{x}_1 = \boldsymbol{x}_2$. If $y_i = j$ for $j > 0$, from the definition of $\underline{\boldsymbol{\alpha}}$-labeling sequences, it holds that $\boldsymbol{x}_{1[i;2]} = \boldsymbol{x}_{2[i;2]} = \boldsymbol{\alpha}_j$. Otherwise, assume $y_i = 0$ and denote by i_ℓ the largest index such that $i_\ell \leq i$, and $y_{i_\ell} \neq 0$. Additionally, let i_r be the smallest index such that $i_r \geq i$, and $y_{i_r} \neq 0$. From the definition of $\underline{\boldsymbol{\alpha}}$-labeling, it holds that $\boldsymbol{x}_{1[i';2]} \in \mathcal{S}$ for $i_\ell \leq i' < i_r$. There is only one path in G of length $m = i_r - i_\ell$ between any two vertices. And so, $\boldsymbol{x}_{1[i_\ell;m]}$ is uniquely determined, so $\boldsymbol{x}_{1[i_\ell;m]} = \boldsymbol{x}_{2[i_\ell;m]}$.

On the other direction, assume that G is not path unique. So, there exist two vertices $u, v \in V$ with two different paths between u and v of the same length $m > 1$. Denote these two paths by $w_1 = ut_1t_2 \cdots t_{m-1}v$ and $w_2 = us_1s_2 \cdots s_{m-1}v$ when $t_i, s_i \in V$ for $1 \leq i \leq m-1$, $(t_i, t_{i+1}), (s_i, s_{i+1}) \in E$ for $1 \leq i \leq m-2$, $(u, t_1), (t_{m-1}, v), (u, s_1), (s_{m-1}, v) \in E$. From the definition of E, $t_it_{i+1}, s_is_{i+1} \in \mathcal{S}$ for $1 \leq i \leq m-2$, and $ut_1, t_{m-1}v, us_1, s_{m-1}v \in \mathcal{S}$. So, $L_{\underline{\boldsymbol{\alpha}}}(w_1) = L_{\underline{\boldsymbol{\alpha}}}(w_2) = (0, \ldots, 0)$. In other words, the function $L_{\underline{\boldsymbol{\alpha}}}(\cdot)$ does not distinguish between the substrings w_1 and w_2.

Denote the set of all sequences of length n over Σ_4 which do not contain w_2 as a substring by $\mathcal{K}_{w_2} = \{\boldsymbol{w} \in \Sigma_4^n | \boldsymbol{w} \neq \boldsymbol{p}w_2\boldsymbol{q}, \boldsymbol{p}, \boldsymbol{q} \in \Sigma_4^*\}$. Let $L_{\underline{\boldsymbol{\alpha}}}^* : \mathcal{K}_{w_2} \to \Sigma_{|\mathcal{S}|+1}^n$ be a function for which $\forall \boldsymbol{w} \in \mathcal{K}_{w_2}, L_{\underline{\boldsymbol{\alpha}}}^*(\boldsymbol{w}) = L_{\underline{\boldsymbol{\alpha}}}(\boldsymbol{w})$. Let $\boldsymbol{w}' \in \Sigma_4^n \setminus \mathcal{K}_{w_2}$ be a sequence that contains w_2 as a substring. Let $\boldsymbol{w}^* \in \Sigma_4^n$ be such that $\boldsymbol{w}^*_{[i;m+1]} = w_1$ if $\boldsymbol{w}'_{[i;m+1]} = w_2$ and $w_i^* = w_i'$ otherwise. It holds that $\boldsymbol{w}^* \in \mathcal{K}_{w_2}$ and so $L_{\underline{\boldsymbol{\alpha}}}(\boldsymbol{w}') = L_{\underline{\boldsymbol{\alpha}}}(\boldsymbol{w}^*) = L_{\underline{\boldsymbol{\alpha}}}^*(\boldsymbol{w}^*)$. As a result, $|\mathrm{Im}(L_{\underline{\boldsymbol{\alpha}}})| = |\mathrm{Im}(L_{\underline{\boldsymbol{\alpha}}}^*)| \leq |\mathcal{K}_{w_2}|$, where $\mathrm{Im}(f)$ is the image of f. Since the length of w_2 is fixed, it holds that $\mathsf{cap}(\mathcal{K}_{w_2}) < 2$, and thus $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) < 2$. ∎

The proof of Theorem 7 will be divided into two claims.

Claim 1. There exist ten labels of length two, $\underline{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{10})$, such that $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = 2$.

Proof: Consider the set of the following six labels: $\mathcal{S} = \{AA, AC, AT, GG, GC, GT\}$ and $\underline{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{10})$ when $\forall i \in [10], \boldsymbol{\alpha}_i \in \bar{\mathcal{S}}$. Let $G = (V, E)$ be a directed graph in which $V = \Sigma_4$ and $E = \{(x, y) | xy \in \mathcal{S}\}$. Since G is a path unique graph, from Theorem 8, $\mathsf{cap}(\underline{\boldsymbol{\alpha}}) = 2$.

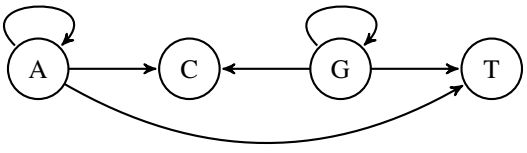


Fig. 5: The Constrained Graph G for Claim 1. ∎

Claim 2. Let $G = (V, E)$ be a path unique graph with $V = \Sigma_4$. So, $|E| \leq 6$.

The proof is based on dividing into cases according to the number of self-loops in the graph.

REFERENCES

[1] A. Moter and U. B. Göbel, "Fluorescence in situ hybridization (FISH) for direct visualization of microorganisms," *Journal of microbiological methods*, vol. 41, no. 2, pp. 85–112, 2000.

[2] B. Chen, W. Zou, H. Xu, Y. Liang, and B. Huang, "Efficient labeling and imaging of protein-coding genes in living cells using CRISPR-tag," *Nature communications*, vol. 9, no. 1, p. 5065, 2018.

[3] D. Gruszka, J. Jeffet, S. Margalit, Y. Michaeli, and Y. Ebenstein, "Single-molecule optical genome mapping in nanochannels: Multidisciplinarity at the nanoscale," *Essays in Biochemistry*, vol. 65, no. 1, pp. 51–66, 2021.

[4] H. Ma, A. Naseri, P. Reyes-Gutierrez, S. A. Wolfe, S. Zhang, and T. Pederson, "Multicolor crispr labeling of chromosomal loci in human cells," *Proceedings of the National Academy of Sciences*, vol. 112, no. 10, pp. 3002–3007, 2015.

[5] J. Deen, C. Vranken, V. Leen, R. K. Neely, K. P. Janssen, and J. Hofkens, "Methyltransferase-directed labeling of biomolecules and its applications," *Angewandte Chemie International Edition*, vol. 56, no. 19, pp. 5182–5200, 2017.

[6] S. Ohayon, A. Girsault, M. Nasser, S. Shen-Orr, and A. Meller, "Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification," *PLoS computational biology*, vol. 15, no. 5, p. e1007067, 2019.

[7] J. A. Alfaro, P. Bohländer, M. Dai, M. Filius, C. J. Howard, X. F. Van Kooten, S. Ohayon, A. Pomorski, S. Schmid, A. Aksimentiev *et al.*, "The emerging landscape of single-molecule protein sequencing technologies," *Nature methods*, vol. 18, no. 6, pp. 604–617, 2021.

[8] A. P. Young, D. J. Jackson, and R. C. Wyeth, "A technical review and guide to rna fluorescence in situ hybridization," *PeerJ*, vol. 8, p. e8806, 2020.

[9] B. Canard and R. S. Sarfati, "Dna polymerase fluorescent substrates with reversible 3-tags," *Gene*, vol. 148, no. 1, pp. 1–6, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0378111994902267

[10] M. F. Fraga and M. Esteller, "Dna methylation: A profile of methods and applications," *BioTechniques*, vol. 33, no. 3, pp. 632–649, 2002, pMID: 12238773. [Online]. Available: https://doi.org/10.2144/02333rv01

[11] H. Frickmann, A. E. Zautner, A. Moter, J. Kikhney, R. M. Hagen, H. Stender, and S. Poppert, "Fluorescence in situ hybridization (fish) in the microbiological diagnostic routine laboratory: a review," *Critical Reviews in Microbiology*, vol. 43, no. 3, pp. 263–293, 2017, pMID: 28129707. [Online]. Available: https://doi.org/10.3109/1040841X.2016.1169990

[12] M. Levy-Sakin and Y. Ebenstein, "Beyond sequencing: Optical mapping of DNA in the age of nanotechnology and nanoscopy," *Current opinion in biotechnology*, vol. 24, no. 4, pp. 690–698, 2013.

[13] V. Müller and F. Westerlund, "Optical DNA mapping in nanofluidic devices: Principles and applications," *Lab on a Chip*, vol. 17, no. 4, pp. 579–590, 2017.

[14] B. Dey, S. Thukral, S. Krishnan, M. Chakrobarty, S. Gupta, C. Manghani, and V. Rani, "DNA-protein interactions: Methods for detection and analysis," *Molecular and cellular biochemistry*, vol. 365, pp. 279–99, 03 2012.

[15] B. H. Marcus, R. M. Roth, and P. H. Siegel, "An introduction to coding for constrained systems," *Lecture notes*, 2001.

[16] S. McLaughlin, J. Luo, and Q. Xie, "On the capacity of m-ary runlength-limited codes," *IEEE Transactions on Information Theory*, vol. 41, no. 5, pp. 1508–1511, 1995.