On the Number of DNA Sequence Profiles for Practical Values of Read Lengths

Zuling Chang*, Johan Chrisnata†, Martianus Frederic Ezerman†, and Han Mao Kiah†
*School of Mathematics and Statistics, Zhengzhou University, China
†School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
Emails: zuling_chang@zzu.edu.cn, {jchrisnata, fredezerman, hmkiah}@ntu.edu.sg

***Abstract*—A recent study by one of the authors has demonstrated the relevance of profile vectors in DNA-based data storage. We provide exact values and lower bounds on the number of profile vectors for finite values of alphabet size q, read length ℓ, and word length n. Consequently, we demonstrate that for $q \geq 3$ and $n = q^a\ell$, $a = o(\ell)$, the number of profile vectors is at least $q^{\kappa n}$ for some constant $0 < \kappa \leq 1$. In addition to enumeration results, we provide a set of efficient encoding and decoding algorithms for a family of profile vectors.**

***Index Terms*—DNA-based data storage, profile vectors, Lyndon words, synchronization.**

1. Introduction

Despite advances in traditional data recording techniques, the emergence of Big Data platforms and energy conservation issues impose new challenges to the storage community in terms of identifying high volume, nonvolatile, and durable recording media. The potential for using macromolecules for ultra-dense storage was recognized as early as in the 1960s. Among these macromolecules, DNA molecules stand out due to their biochemical robustness and high storage capacity.

In the last few decades, the technologies for synthesizing (writing) artificial DNA and for massive sequencing (reading) have reached unprecedented levels of efficiency and accuracy. Building upon the rapid growth of DNA synthesis and sequencing technologies, two laboratories recently outlined architectures for archival DNA-based storage [1], [2]. The first architecture achieved a density of 700 TB/gram, while the second approach raised the density to 2.2 PB/gram. To further protect against errors, Grass *et al.* later incorporated Reed-Solomon error-correction schemes and encapsulated the DNA media in silica [3]. Yazdi *et al.* recently proposed a completely different approach and provided a random access and rewritable DNA-based storage system [4], [5].

More recently, to control specialized errors arising from sequencing platforms, two families of codes were introduced by Gabrys *et al.* [6] and Kiah *et al.* [7]. The former looks at miniaturized nanopore sequencers such as MinION, while the latter focuses on errors arising from high-throughput sequencers such as Illumina. The latter forms the basis for this work. In particular, we examine the concept of *DNA profile vectors* introduced by Kiah *et al.* [7].

In this channel model, to store and retrieve information in DNA, one starts with a desired information sequence encoded into a sequence defined over the nucleotide alphabet $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$. The *DNA storage channel* models a physical process which takes as its input the sequence of length n, and synthesizes (writes) it physically into a macromolecule string. To retrieve the information, the user may proceed using several read technologies. The most common sequencing process, implemented by Illumina, makes numerous copies of the string or amplifies the string, and then fragments all copies of the string into a collection of substrings (reads) of approximately the same length ℓ, so as to produce a large number of overlapping "reads". Since the concentration of all (not necessarily) distinct substrings within the mix is usually assumed to be uniform, one may normalize the concentration of all subsequences by the concentration of the least abundant substring. As a result, one actually observes substring concentrations reflecting the frequency of the substrings in *one copy* of the original string. Therefore, we model the output of the channel as an *unordered subset of substrings (reads)*, and this set may be summarized by its multiplicity vector, which we call the *output profile vector*.

We assume a channel with neither synthesis nor sequencing errors, and observe that it is possible for different strings to have an identical profile vector. In other words, even without errors, the channel may be unable to distinguish between certain pairs of strings. Our task is then to enumerate all distinct profile vectors for fixed values of n and ℓ over a q-ary alphabet. In the case of arbitrary ℓ-substrings, the problem of enumerating all valid profile vectors was addressed by Jacquet *et al.* in the context of "Markov types" [8]. Kiah *et al.* then extended the enumeration results to profiles with specific ℓ-substring constraints so as to address certain considerations in DNA sequence design [7]. In particular, for fixed values of q and ℓ, the number of profile vectors is known to be $\Theta\left(n^{q^\ell - q^{\ell-1}}\right)$.

However, determining the coefficient for the dominating term $n^{q^\ell - q^{\ell-1}}$ is a computationally difficult task. It has been determined for only very small values of q and ℓ in [7], [8]. Furthermore, it is unclear how accurate the asymptotic estimate $\Theta\left(n^{q^\ell - q^{\ell-1}}\right)$ is for practical values of n. Indeed, most current DNA storage systems do not use string lengths n exceeding several thousands nucleotides (nts) due to the high cost of synthesis. On the other hand, current sequencing systems have read length ℓ between 100 to 1500 nts.

In this paper, we adopt a different approach and look for lower bounds for the number of profile vectors given moderate values of q, ℓ, and n. Surprisingly, for fixed $q \geq 3$ and moderately large values $n = q^a\ell$ with $a = o(\ell)$, the number of profile vectors is at least $q^{\kappa n}$ for some constant $0 < \kappa \leq 1$. As an example, when $q = 4$ (the number of DNA nucleotide bases) and $\ell = 100$ (a practical read length), we show that there are at least $4^{0.753n}$ distinct profile vectors for $n \leq 25600$. In other words, for practical values of read and word lengths, we are able to obtain a set of distinct profile vectors with strictly positive rates.

In addition to enumeration results, we demonstrate a set of linear-time encoding and decoding algorithms for a family of profile vectors.

2. Preliminaries

Let $[\![q]\!]$ denote the set of integers $\{0, 1, \ldots, q-1\}$ and consider a word $\mathbf{x} = x_1x_2\cdots x_n$ of length n over $[\![q]\!]$. For $1 \leq i < j \leq n$, we denote the entry x_i by $\mathbf{x}[i]$, the *substring* $x_ix_{i+1}\cdots x_j$ of length $(j-i+1)$ by $\mathbf{x}[i,j]$, and the length of $\mathbf{x}$ by $|\mathbf{x}|$.

For $\ell \leq n$ and $1 \leq i \leq n-\ell+1$, we also call the substring $\mathbf{x}[i, i+\ell-1]$ an *ℓ-gram* of $\mathbf{x}$. For $\mathbf{z} \in [\![q]\!]^\ell$, let $p(\mathbf{x}, \mathbf{z})$ denote the number of occurrences of $\mathbf{z}$ as an ℓ-gram of $\mathbf{x}$. Let $\mathbf{p}(\mathbf{x}, \ell) \triangleq \Big(p(\mathbf{x}, \mathbf{z})\Big)_{\mathbf{z}\in[\![q]\!]^\ell}$ be the (ℓ-gram) *profile vector* of length q^ℓ, indexed by all words of $[\![q]\!]^\ell$ ordered lexicographically. Let $\mathcal{F}(\mathbf{x}, \ell)$ be the set of ℓ-grams of $\mathbf{x}$. In other words, $\mathcal{F}(\mathbf{x}, \ell)$ is the support for the vector $\mathbf{p}(\mathbf{x}, \ell)$.

Example 2.1. Let $q = 2$, $n = 5$ and $\ell = 2$. Then $p(10001, 01) = p(10001, 10) = 1$, while $p(10001, 00) = 2$. So, $\mathbf{p}(10001, 2) = (2, 1, 1, 0)$ and $\mathcal{F}(10001, 2) = \{00, 01, 10\}$.

Consider the words 00010 and 00101. Then $\mathbf{p}(10001, 2) = \mathbf{p}(00010, 2)$ while $\mathcal{F}(10001, 2) = \mathcal{F}(00010, 2) = \mathcal{F}(00101, 2)$.

As illustrated by Example 2.1, different words may have the same profile vector. We define a relation on $[\![q]\!]^n$ where $\mathbf{x} \sim \mathbf{x}'$ if and only if $\mathbf{p}(\mathbf{x}, \ell) = \mathbf{p}(\mathbf{x}', \ell)$. It can be shown that $\sim$ is an equivalence relation and we denote the number of equivalence classes by $P_q(n, \ell)$. We further define the *rate of profile vectors* to be $R_q(n, \ell) = \log_q P_q(n, \ell)/n$. The asymptotic growth of $P_q(n, \ell)$ as a function of n is given as below.

Theorem 2.1 (Jacquet *et al.* [8], Kiah *et al.* [7]). Fix $q \geq 2$ and ℓ. Then

$$P_q(n, \ell) = \Theta\left(n^{q^\ell - q^{\ell-1}}\right).$$

Hence, $\lim_{n\to\infty} R_q(n, \ell) = 0$.

Our main contribution is the following set of lower bounds for $P_q(n, \ell)$ for finite values of n, q and ℓ.

Theorem 2.2. Fix $q \geq 2$ and $n \geq \ell$,

(i) If $\ell \leq n < 2\ell$, then

$$P_q(n, \ell) = q^n - \sum_{r|n-\ell+1} \sum_{t|r} \left(\frac{r-1}{r}\right) \mu\left(\frac{r}{t}\right) q^t, \quad (1)$$

where μ is the Möbius function.

(ii) If $n = q^{a-1}\ell$ where $4 \leq 2a \leq \ell$, then

$$P_q(n, \ell) \geq (q-1)^{q^{a-1}(\ell-a)}. \quad (2)$$

We prove Equations (1) and (2) in Sections 3 and 4, respectively.

Example 2.2. Setting $q = 4$, $a = 5$, $\ell = 100$ in (2) yields $P_4(25600, 100) \geq 3^{24320} \approx 4^{19273}$. In other words, $R_4(25600, 100) \geq 0.753$. While (2) is stated for words of length $n = q^{a-1}\ell$, we modify the construction to obtain words of length n where $q^{a-2}\ell < n < q^{a-1}\ell$ for some a and when ℓ divides n. The details are given at the end of Section 4. Hence, by varying $a \in \{2, 3, 4, 5\}$ in (2), we have

$$R_4(100n', 100) \geq \begin{cases} 0.753, & \text{for } 64 < n' \leq 256; \\ 0.761, & \text{for } 16 < n' \leq 64; \\ 0.768, & \text{for } 4 < n' \leq 16; \\ 0.777, & \text{for } 1 < n' \leq 4. \end{cases}$$

Furthermore, from (1), we can compute that $R_4(n, 100) \approx 1$ for $100 \leq n < 200$.

We now provide an asymptotic analysis for the rates of profile vectors. Let n be a function of ℓ, or $n = n(\ell)$ such that $n(\ell)$ increases with ℓ. We then define the *asymptotic rate of profile vectors with respect to n* via the equation

$$\alpha(n, q) \triangleq \lim_{\ell\to\infty} R_q(n, \ell). \quad (3)$$

Suppose that ℓ is a system parameter determined by current sequencing technology. Then $n = n(\ell)$ determines how long we can set our codewords so that the information rate of the DNA storage channel remains as $\alpha(n, q)$.

From Theorem 2.2, we derive the following results on the asymptotic rates.

Corollary 2.3 (Asymptotic rates)**.** Fix $q \geq 2$.

(i) If $n = \lfloor \lambda\ell \rfloor$ for some constant $1 \leq \lambda < 2$, then

$$\alpha(n, q) = 1. \quad (4)$$

(ii) If $n = q^a\ell$ with $a = o(\ell)$, then

$$\alpha(n, q) \geq \log_q(q-1). \quad (5)$$

3. Exact Enumeration of Profile Vectors

We extend the methods of Tan and Shallit [9], where the number of possible $\mathcal{F}(\mathbf{x}, \ell)$ was determined for $\ell \leq n < 2\ell$. Specifically, we compute $P_q(n, \ell)$ for $\ell \leq n < 2\ell$. Our strategy is to first define an equivalence relation using the notions of root conjugates so that the number of equivalence classes yields $P_q(n, \ell)$. We then compute this number using standard combinatorial methods.

Definition 3.1. Let $\mathbf{x}$ be a q-ary word. A *period* of $\mathbf{x}$ is a positive integer r such that $\mathbf{x}$ can be *factorized* as

$$\mathbf{x} = \underbrace{\mathbf{u}\,\mathbf{u}\cdots\mathbf{u}}_{k \text{ times}}\mathbf{u}', \text{ with } |\mathbf{u}| = r, \mathbf{u}' \text{ a prefix of } \mathbf{u}, \text{ and } k \geq 1.$$

Let $\pi(\mathbf{x})$ denote the *minimum period* of $\mathbf{x}$. The *root* of $\mathbf{x}$ is given by $\mathbf{r}(\mathbf{x}) = \mathbf{x}[1, \pi(\mathbf{x})]$, which is the prefix of $\mathbf{x}$ with length $\pi(\mathbf{x})$. Two words $\mathbf{x}$ and $\mathbf{x}'$ are said to be *root-conjugate* if $\mathbf{r}(\mathbf{x}) = \mathbf{uv}$ and $\mathbf{r}(\mathbf{x}') = \mathbf{vu}$ for some words $\mathbf{u}$ and $\mathbf{v}$, or $\mathbf{r}(\mathbf{x})$ is a *rotation* of $\mathbf{r}(\mathbf{x}')$.

Example 3.1. 10010010 has minimal period three and its root is 100. Also, 01001001 has minimal period three and its root is 010. Therefore, 10010010 and 01001001 are root-conjugates.

Observe that two words that are root-conjugates necessarily have the same minimal period and it can be shown that being root-conjugates form an equivalence relation. In addition, we have the following technical lemma.

Lemma 3.1. Let $\mathbf{x}$ be a word of length n with $\pi(\mathbf{x}) \le n-\ell+1 \le \ell$. Then for $1 \le i < j \le \pi(\mathbf{x})$, we have $\mathbf{x}[i, i+\ell-1] \ne \mathbf{x}[j, j+\ell-1]$.

Proof. Suppose that $\mathbf{x}[i, i+\ell-1] = \mathbf{x}[j, j+\ell-1]$. Letting $k = j - i$, we have $\mathbf{x}[s] = \mathbf{x}[s+k]$ for $i \le s \le i+\ell-1$. Since $\pi(\mathbf{x}) \le n-\ell+1 \le \ell$, then $\mathbf{x}[s] = \mathbf{x}[s+k]$ for $1 \le s \le \pi(\mathbf{x})$. Therefore, $\mathbf{x}[s] = \mathbf{x}[s+d]$ for $1 \le s \le \pi(\mathbf{x})$ where $d = \gcd(k, \pi(\mathbf{x})) \le k = j-i < \pi(\mathbf{x})$. In other words, $\mathbf{x}$ has a period $d < \pi(\mathbf{x})$, contradicting the minimality of $\pi(\mathbf{x})$. $\square$

Tan and Shallit proved the following result that characterized $\mathcal{F}(\mathbf{x}, \ell)$ when $|\mathbf{x}| < 2\ell$.

Lemma 3.2 (Tan and Shallit [9, Th. 15]). Suppose that $\ell \le n < 2\ell$ and $\mathbf{x}$ and $\mathbf{x}'$ are distinct q-ary words of length n. Then $\mathcal{F}(\mathbf{x}, \ell) = \mathcal{F}(\mathbf{x}', \ell)$ if and only if $\mathbf{x}, \mathbf{x}'$ are root-conjugates with $\pi(\mathbf{x}) \le n-\ell+1$.

Using Lemma 3.1, we extend Lemma 3.2 to characterize the profile vectors when $n < 2\ell$.

Theorem 3.3. Let $\mathbf{x}$ and $\mathbf{x}'$ be distinct q-ary words of length n. If $\mathbf{x}, \mathbf{x}'$ are root-conjugates with $\pi(\mathbf{x}) \mid n-\ell+1$, then $\mathbf{p}(\mathbf{x}, \ell) = \mathbf{p}(\mathbf{x}', \ell)$. Conversely, if $\ell \le n < 2\ell$ and $\mathbf{p}(\mathbf{x}, \ell) = \mathbf{p}(\mathbf{x}', \ell)$, then $\mathbf{x}, \mathbf{x}'$ are root-conjugates with $\pi(\mathbf{x}) \mid n-\ell+1$.

Proof. Suppose that $\mathbf{x}$ and $\mathbf{x}'$ are root-conjugates with $\pi(\mathbf{x}) = r$ and $n-\ell+1 = rs$ for some s. Then it can be verified that $\mathcal{F}(\mathbf{x}, \ell) = \mathcal{F}(\mathbf{x}', \ell) = \{\mathbf{x}[i, i+\ell-1] : 1 \le i \le r\}$ and $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}', \mathbf{z}) = s$ for all $\mathbf{z} \in \mathcal{F}(\mathbf{x}, \ell)$. Therefore, $\mathbf{p}(\mathbf{x}, \ell) = \mathbf{p}(\mathbf{x}', \ell)$.

Conversely, let $\mathbf{p}(\mathbf{x}, \ell) = \mathbf{p}(\mathbf{x}', \ell)$. Then $\mathcal{F}(\mathbf{x}, \ell) = \mathcal{F}(\mathbf{x}', \ell)$. By Lemma 3.2, we have that $\mathbf{x}, \mathbf{x}'$ are root-conjugates with $\pi(\mathbf{x}) \le n-\ell+1$. Let $r = \pi(\mathbf{x})$. It remains to show that $r \mid n-\ell+1$.

Suppose otherwise and let $n-\ell+1 = rs+t$ with $1 \le t \le r-1$. Let the roots of $\mathbf{x}$ and $\mathbf{x}'$ be $\mathbf{uv}$ and $\mathbf{vu}$, respectively. Therefore, we can write $\mathbf{x}$ and $\mathbf{x}'$ as

$$\mathbf{x} = \overbrace{\underbrace{\mathbf{uv}}_{r}\,\underbrace{\mathbf{uv}}_{r}\cdots\underbrace{\mathbf{uv}}_{r}}^{s \text{ times}}\,\underbrace{\mathbf{w}}_{t+\ell-1}$$

$$\mathbf{x}' = \overbrace{\underbrace{\mathbf{vu}}_{r}\,\underbrace{\mathbf{vu}}_{r}\cdots\underbrace{\mathbf{vu}}_{r}}^{s \text{ times}}\,\underbrace{\mathbf{w}'}_{t+\ell-1}$$

We have the following cases.

(i) If $1 \le t < |\mathbf{u}|$, let $\mathbf{z}'$ be the ℓ-length prefix of $\mathbf{x}'$. Since $|\mathbf{w}'| = t+\ell-1 \ge \ell$ and $\mathbf{z}'$ is a prefix of $\mathbf{w}'$, we have $p(\mathbf{x}', \mathbf{z}') \ge s+1$. On the other hand, from Lemma 3.1, the ℓ-gram of $\mathbf{z}'$ can only appear after the first $|\mathbf{u}|$ coordinates of $\mathbf{w}$. However, $|\mathbf{w}|-|\mathbf{u}| < t+(\ell-1)-t < \ell$, and so, there is no occurrence of $\mathbf{z}'$ as an ℓ-gram of $\mathbf{w}$. Therefore, $p(\mathbf{x}, \mathbf{z}') = s < p(\mathbf{x}', \mathbf{z}')$, contradicting the assumption that $\mathbf{p}(\mathbf{x}, \ell) = \mathbf{p}(\mathbf{x}', \ell)$.

(ii) If $|\mathbf{u}| \le t \le r-1$, let $\mathbf{z} = \mathbf{x}[|\mathbf{u}|, |\mathbf{u}|+\ell-1]$. Since $|\mathbf{w}| = t+\ell-1 \ge |\mathbf{u}|+\ell-1$, we have $p(\mathbf{x}, \mathbf{z}) \ge s+1$. With the same considerations as before, we check that there is no occurence of $\mathbf{z}$ as an ℓ-gram of $\mathbf{w}'$, So, $p(\mathbf{x}', \mathbf{z}) = s < p(\mathbf{x}, \mathbf{z})$, a contradiction.

Therefore, we conclude $t = 0$ or $r \mid n-\ell+1$ as desired. $\square$

Hence, for $\ell \le n < 2\ell$, we have $\mathbf{x} \sim \mathbf{x}'$ if and only if $\mathbf{x}$ and $\mathbf{x}'$ are root-conjugates with $\pi(\mathbf{x}) \mid n-\ell+1$ and we compute the number of equivalence classes using this characterization.

A word is said to be *aperiodic* if it is not equal to any of its nontrivial rotations. An aperiodic word of length r is said to be *Lyndon* if it is the lexicographically least word amongst all its r rotations. The number of Lyndon words [10] of length r is given by $L(r) = \left(\sum_{t|r} \mu(r/t) q^t\right)/r$.

For any integer $r \mid n-\ell+1$ and any word $\mathbf{x}$, if $\pi(\mathbf{x}) = r$ and $\mathbf{r}(\mathbf{x})$ is its root, then $\mathbf{r}(\mathbf{x})$ is aperiodic and is a rotation of some Lyndon word $\mathbf{u}(\mathbf{x})$. Let $\mathbf{u}(\mathbf{x})$ be the representative of the equivalence class of $\mathbf{x}$. Since there are r rotations of $\mathbf{u}(\mathbf{x})$, we observe that there are r words in the equivalence class of $\mathbf{x}$. Therefore, we can compute the number of equivalence classes to be

$$q^n - \sum_{r|n-\ell+1} (r-1) L(r),$$

and, consequently, obtain (1).

From Theorem 3.3, if $\mathbf{x}$ and $\mathbf{x}'$ are root-conjugates with $\pi(\mathbf{x}) \mid n-\ell+1$, we have $\mathbf{p}(\mathbf{x}, \ell) = \mathbf{p}(\mathbf{x}', \ell)$ for all values of n. In other words, the number of equivalence classes computed above provides an upper bound for the number of profile vectors. Formally, we have the following corollary.

Corollary 3.4. For $n \ge 2\ell$,

$$P_q(n, \ell) \le q^n - \sum_{r|n-\ell+1} \sum_{t|r} \left(\frac{r-1}{r}\right) \mu\left(\frac{r}{t}\right) q^t.$$

4. Distinct Profile Vectors from Addressable Codes

Borrowing ideas from synchronization, we construct a set of words with different profile vectors and prove (2). Here, our strategy is to mimic the concept of *watermark* and *marker codes* [11]–[13], where a 'marker' pattern is distributed throughout a codeword. Due to the unordered nature of the short reads, instead of a single 'marker' pattern, we consider a set of patterns.

More formally, suppose that $2a \le \ell \le n$. Let $\mathcal{A} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_M\} \subseteq [\![q]\!]^a$ be a set of M sequences of length a. Elements of $\mathcal{A}$ are called *addresses*. A word $\mathbf{x} = \mathbf{z}_1\mathbf{z}_2\cdots\mathbf{z}_M$, where $|\mathbf{z}_i| = \ell$ for all $1 \le i \le M$, is said to be $(\mathcal{A}, \ell)$-*addressable* if the following properties hold.

(C1) The prefix of length a of $\mathbf{z}_i$ is equal to $\mathbf{u}_i$ for all $1 \le i \le M$. In other words, $\mathbf{z}_i[1, a] = \mathbf{u}_i$.

(C2) $\mathbf{z}_i[j, j + a - 1] \notin \mathcal{A}$ for all $1 \le i \le M$ and $2 \le j \le \ell - a + 1$.

Conditions (C1) and (C2) imply that the address $\mathbf{u}_i \in \mathcal{A}$ appears exactly once as the prefix of $\mathbf{z}_i$ and does not appear as an a-gram of any substring $\mathbf{z}_j$ with $j \ne i$. A code $\mathcal{C}$ is $(\mathcal{A}, \ell)$-*addressable* if all words in $\mathcal{C}$ are $(\mathcal{A}, \ell)$-addressable.

Intuitively, given an $(\mathcal{A}, \ell)$-addressable word $\mathbf{x}$, we can make use of the addresses in $\mathcal{A}$ to identify the position of each ℓ-gram in $\mathbf{x}$ and hence, reconstruct $\mathbf{x}$. We formalize this idea in the following theorem.

Theorem 4.1. Let $\mathcal{A}$ be a set of addresses of length a and $2a \le \ell$. Suppose that $\mathcal{C}$ is an $(\mathcal{A}, \ell)$-addressable code. For distinct words $\mathbf{x}, \mathbf{x}' \in \mathcal{C}$, we have $\mathcal{F}(\mathbf{x}, \ell) \ne \mathcal{F}(\mathbf{x}', \ell)$. Therefore, $\mathbf{p}(\mathbf{x}, \ell) \ne \mathbf{p}(\mathbf{x}', \ell)$ and $P_q(n, \ell) \ge |\mathcal{C}|$.

Proof. Let $\mathbf{x} = \mathbf{z}_1\mathbf{z}_2\cdots\mathbf{z}_M$ and $\mathbf{x}' = \mathbf{z}'_1\mathbf{z}'_2\cdots\mathbf{z}'_M$ be distinct $(\mathcal{A}, \ell)$-addressable words in $\mathcal{C}$. Without loss of generality, we assume $\mathbf{z}_1 \ne \mathbf{z}'_1$. Observe that $\mathbf{z}_1 \in \mathcal{F}(\mathbf{x}, \ell)$. To prove the theorem, it suffices to show that $\mathbf{z}_1 \notin \mathcal{F}(\mathbf{x}', \ell)$.

Suppose otherwise that $\mathbf{z}_1$ appears as an ℓ-gram in $\mathbf{x}'$. Since $\mathbf{u}_1$ is a prefix of $\mathbf{z}_1$ with $\mathbf{z}_1 \ne \mathbf{z}'_1$, by Conditions (C1) and (C2), we have that

$$\mathbf{x}' = \cdots \overbrace{\circ\circ\underbrace{\oplus\oplus\cdots\oplus}_{|\mathbf{u}_i|=a}\oplus\oplus}^{|\mathbf{z}_1|=\ell}++\cdots \quad \text{for some } i \ne 1.$$

Here, $\circ$'s and $+$'s represent the ℓ-grams $\mathbf{z}_1$ and $\mathbf{z}'_i$, respectively, and $\oplus$'s indicate the symbols that are in the overlap of the two ℓ-grams. Since $2a \le \ell$, $\mathbf{u}_i$ must be in $\mathbf{z}_1$ as an a-gram, contradicting Condition (C2). $\square$

To employ Theorem 4.1, we define the following set of addresses,

$$\mathcal{A}^* \triangleq \left\{(u_1, u_2, \dots, u_a) : \sum_{i=1}^{a} u_i = 0 \bmod q\right\}. \tag{6}$$

So, $\mathcal{A}^*$ is a set of $M = q^{a-1}$ addresses and we list the addresses as $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_M$. To construct an $(\mathcal{A}^*, \ell)$-addressable code, we consider the encoding map encode : $\{1, 2, \dots, q-1\}^{(\ell-a)M} \to [\![q]\!]^{M\ell}$ given in Algorithm 1 and define $\mathcal{C}$ to be the image of encode. Conversely, we consider the decoding map decode : $\mathcal{C} \to \{1, 2, \dots, q-1\}^{(\ell-a)M}$ given in Algorithm 2.

Example 4.1. For $q = 4$, $a = 2$, $\mathcal{A}^* = \{00, 13, 22, 31\}$ by (6). Consider $\ell = 5$ and the data string $\mathbf{c} = (111, 123, 222, 321)$. Applying Algorithm 1 to construct $\mathbf{z}_1$ with $\mathbf{c}_1 = 111$, we start with $\mathbf{z}_1 = 00$. Then $z_{\text{bad}} = 0$ and we choose the first element of $\{1, 2, 3\}$ to augment $\mathbf{z}_1$ to 001. In the next iteration, we have $z_{\text{bad}} = 3$ and augment $\mathbf{z}_1$ to 0010. Repeating this, we then obtain $\mathbf{z}_1 = \underline{00}101$. More generally, we have

$$\mathbf{z}_1 = \underline{00}101,\ \mathbf{z}_2 = \underline{13}023,\ \mathbf{z}_3 = \underline{22}111,\ \mathbf{z}_4 = \underline{31}210,$$

and so, encode$(\mathbf{c}) = (00101, 13023, 22111, 31210) = \mathbf{x}$. We check that $\mathbf{x}$ is indeed $(\mathcal{A}^*, \ell)$-addressable.

We also verify that decode$(\mathbf{x})$ in Algorithm 2 indeed returns the data string $\mathbf{c}$. Since there are 3^{12} possible data strings, $|\mathcal{C}| = 3^{12} \approx 4^{9.51}$.

Algorithm 1 encode$(\mathbf{c}, \mathcal{A}^*)$

Input: Data string $\mathbf{c} = \mathbf{c}_1\mathbf{c}_2\cdots\mathbf{c}_M$,
where $\mathbf{c}_i \in \{1, 2, \dots, q-1\}^{(\ell-a)}$ for $1 \le i \le M$,
and $\mathcal{A}^*$ is defined by (6).

Output: $\mathbf{x} = \mathbf{z}_1\mathbf{z}_2\cdots\mathbf{z}_M \in [\![q]\!]^{M\ell}$, where $\mathbf{x}$ is $(\mathcal{A}^*, \ell)$-addressable.

for $1 \le i \le M$ **do**
  $\mathbf{z}_i \leftarrow \mathbf{u}_i$ ($\mathbf{u}_i$ has length a)
  for $a + 1 \le j \le \ell$ **do**
    $z_{\text{bad}} \leftarrow -\sum_{s=1}^{a-1} \mathbf{z}_i[j - s] \bmod q$
      (sum of the last $a - 1$ entries modulo q)
    $z \leftarrow \mathbf{c}_i[j - a]$-th element of $([\![q]\!] \setminus \{z_{\text{bad}}\})$
    append $\mathbf{z}_i$ with z
  end for
end for
return $\mathbf{z}_1\mathbf{z}_2\cdots\mathbf{z}_M$

Algorithm 2 decode$(\mathbf{z}_1\mathbf{z}_2\cdots\mathbf{z}_M)$

Input: Codeword $\mathbf{z}_1\mathbf{z}_2\cdots\mathbf{z}_M \in \mathcal{C}$.

Output: $\mathbf{c}_1\mathbf{c}_2\cdots\mathbf{c}_M \in \{1, 2, \dots, q-1\}^{(\ell-a)M}$.

for $1 \le i \le M$ **do**
  for $a + 1 \le j \le \ell$ **do**
    $z_{\text{bad}} \leftarrow -\sum_{s=1}^{a-1} \mathbf{z}_i[j - s] \bmod q$
      (sum of the last $a - 1$ entries modulo q)
    $\mathbf{c}_i[j - a] \leftarrow$ the index of the element of $([\![q]\!] \setminus \{z_{\text{bad}}\})$
  end for
end for
return $\mathbf{c}_1\mathbf{c}_2\cdots\mathbf{c}_M$

Algorithm 1 bears similarities with a *linear feedback shift register* [14]. The main difference is that we augment our codeword with a symbol that is *not equal* to the value defined by the linear equation. This then guarantees that we have no a-grams belonging to $\mathcal{A}^*$. More formally, we have the following proposition.

Proposition 4.2. Consider the maps encode, decode and the code $\mathcal{C}$ defined by Algorithms 1 and 2. Then $\mathcal{C}$ is an $(\mathcal{A}^*, \ell)$-addressable code and decode $\circ$ encode$(\mathbf{c}) = \mathbf{c}$ for all $\mathbf{c} \in \{1, 2, \dots, q-1\}^{(\ell-a)M}$. Hence, $|\mathcal{C}| \ge (q-1)^{M(\ell-a)}$. Furthermore, decode and encode computes their respective strings in $O(qM\ell)$ time.

Theorem 4.1 and Proposition 4.2 then yield (2) for $n = q^{a-1}\ell$ and $2a \leq \ell$. In other words, for $n = q^{a-1}\ell$ and $2a \leq \ell$, we have

$$R_q(n,\ell) \geq \left(1 - \frac{a}{\ell}\right)\log_q(q-1).$$

We now modify our construction to derive addressable codes for all values of $n \leq q^{\lfloor \ell/2 \rfloor - 1}\ell$. Suppose that $m = \lfloor n/\ell \rfloor$. Choose $a = \lceil \log_q m \rceil + 1$ so that $m \leq q^{a-1}$. Use a subset $\mathcal{B}^*$ of $\mathcal{A}^*$ of size m for the address set. A straightforward modification of Algorithm 1 then yields $(\mathcal{B}^*,\ell)$-addressable words of the form

$$\mathbf{u}_1 \underbrace{\circ\circ\cdots\circ}_{\ell-a} \mathbf{u}_2 \underbrace{\circ\circ\cdots\circ}_{\ell-a} \cdots \mathbf{u}_m \underbrace{\circ\circ\cdots\circ}_{\ell-a} \underbrace{00\cdots0}_{n-m\ell}.$$

The size of this $(\mathcal{B}^*,\ell)$-addressable code can be computed to be $(q-1)^{m(\ell-a)+\max((t-a),0)}$. We obtain the following corollary.

Corollary 4.3. For $n \leq q^{\lfloor \ell/2 \rfloor - 1}\ell$, suppose that $n = m\ell + t$ with $0 \leq t < \ell$. Set $a = \lceil \log_q m \rceil + 1$ so that $m \leq q^{a-1}$. Then $P_q(n,\ell) \geq (q-1)^{m(\ell-a)}$, or,

$$R_q(n,\ell) \geq \left(\frac{m(\ell-a)}{n}\right)\log_q(q-1) \approx \left(1 - \frac{a}{\ell}\right)\log_q(q-1).$$

5. Conclusion

We adapted ideas from combinatorics of words and synchronizing codes to provide exact values and lower bounds for the number of profile vectors given moderate values of q, ℓ, and n. Surprisingly, for fixed $q \geq 3$ and moderately large values of $n = q^a\ell$ with $a = o(\ell)$, the number of profile vectors is at least $q^{\kappa n}$ for some constant $0 < \kappa \leq 1$. Hence, for practical values of read and word lengths, we are able to obtain a set of distinct profile vectors with strictly positive rates.

In our future work, we want to investigate other functions $n = n(\ell)$ that guarantee a positive asymptotic rate of profile vectors $\alpha(n,q)$ (see (3)) and to examine the number of profile vectors with specific ℓ-gram constraints *a la* Kiah *et al.* [7].

Acknowledgement

The work of Z. Chang is supported by the Joint Fund of the National Natural Science Foundation of China under Grant U1304604. Research Grants TL-9014101684-01 and MOE2013-T2-1-041 support M. F. Ezerman. The authors thank the anonymous reviewers and members of the TPC whose comments improved the presentation of this paper.

References

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

[2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, pp. 77–80, 2013.

[3] 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," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, 2015.

[4] S. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Scientific Reports*, vol. 5, no. 14138, 2015.

[5] S. Yazdi, H. M. Kiah, E. R. Garcia, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *arXiv preprint arXiv:1507.01611*, 2015.

[6] R. Gabrys, H. M. Kiah, and O. Milenkovic, "Asymmetric Lee distance codes for DNA-based storage," *arXiv preprint arXiv:1506.00740*, 2015.

[7] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," *arXiv preprint arXiv:1502.00517*, 2015.

[8] P. Jacquet, C. Knessl, and W. Szpankowski, "Counting Markov types, balanced matrices, and Eulerian graphs," *IEEE Trans. Inform. Theory*, vol. 58, no. 7, pp. 4261–4272, 2012.

[9] S. Tan and J. Shallit, "Sets represented as the length-n factors of a word," in *Combinatorics on Words*. Springer, 2013, pp. 250–261.

[10] R. C. Lyndon, "On Burnside's problem," *Transactions of the American Mathematical Society*, vol. 77, no. 2, pp. 202–215, 1954.

[11] F. Sellers, "Bit loss and gain correction code," *Information Theory, IRE Transactions on*, vol. 8, no. 1, pp. 35–38, 1962.

[12] N. Kashyap and D. L. Neuhoff, "Codes for data synchronization with timing," in *Proc. Data Compression Conference*. IEEE, 1999, pp. 443–452.

[13] M. C. Davey and D. J. MacKay, "Reliable communication over channels with insertions, deletions, and substitutions," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 687–698, 2001.

[14] S. W. Golomb, *Shift register sequences*. Aegean Park Press, 1982.