Codes for DNA Sequence Profiles

Han Mao Kiah*, Gregory J. Puleo†, Olgica Milenkovic‡
*School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore,
†‡Coordinated Science Laboratory, University of Illinois, Urbana-Champaign

***Abstract*—We consider the problem of storing information on synthetic DNA media and associated coding paradigms. The focal question of our analysis it how to construct and enumerate sequences that may be discriminated based on their collection of substrings observed through two types of noisy sequencing channels. In particular, we consider DNA sequences with balanced GC content, needed for chemical stability and desirable hybridization properties. We show that restricted de Bruijn graphs and Ehrhart theory for rational polytopes provide a suitable framework for studying such combinatorial questions.**

1. Introduction

Reconstructing sequences based on partial information about their subsequences, substrings, or composition is an important problem arising in channel synchronization systems, phylogenomics, genomics, and proteomic sequencing [1]–[3]. With the recent development of archival DNA-based storage devices [4], [5] and rewritable, random-access storage media [6], a new family of reconstruction questions has emerged regarding how to *design sequences* which can be easily and accurately reconstructed based on their substrings, in the presence of read and write errors. The write process reduces to DNA synthesis, while the read process involves both DNA sequencing and assembly. The assembly procedure is NP-hard under most formulations [7]. Nevertheless, practical approximation algorithms based on Eulerian paths in de Bruijn graphs have shown to offer good reconstruction performance under the high-coverage model [8].

In our setting, we assume that a sequence $\mathbf{x} \in \{A, T, G, C\}^n$ is first designed according to some combinatorial rules, synthesized[1], and then fragmented into a collection of substrings of approximately the same length ℓ. The latter process models the sequencing strategy used by most modern high-throughput sequencers. The resulting substrings are usually referred to as *reads*. Ideally, one would like to synthesize $\mathbf{x}$ and sequence all ℓ-substrings without errors, which is not possible in practice. In addition to symbol substitution errors occurring both during synthesis and sequencing, a number of substrings may be unavailable for sequencing, leaving coverage gaps in the original message.

To model this read-write phenomena, we introduced in our companion paper [10] the notion of a *DNA storage channel* that takes as its input a sequence $\mathbf{x}$ of length n, introduces substitution errors in $\mathbf{x}$, with the resulting sequence denoted by $\widetilde{\mathbf{x}}$. The channel proceeds to output all or a subset of substrings of the sequence $\widetilde{\mathbf{x}}$ of length ℓ, $\ell < n$. Each of the substrings is allowed to have additional substitution errors, due to sequencing, and some substrings may be missing. The substrings at the output of the DNA storage channel are collectively enumerated by a vector $\widehat{\mathbf{x}}$, termed the channel output (see Fig. 1 for an illustration). In [10], we also introduced a new family of codes capable of correcting synthesis, lack of coverage and sequencing errors arising in the DNA storage channel, all of which may be characterized by asymmetric errors studied in classical coding theory. In addition, we design codebooks whose codewords have different substring counts or substring counts at a "sufficiently large" distance from each other.

We continue our study of this new coding paradigm by *modeling the read process (sequencing)* through the use of *profile vectors*. A profile vector of a sequence enumerates all substrings of the sequence, and we enumerate the equivalence classes under the channel mapping. In addition, we design new coding techniques that make use of codewords with ℓ-substrings of high biochemical stability which are also resilient to errors. For this purpose, we consider a number of *codeword constraints* known to influence the performance of both the synthesis and sequencing systems, one of which we termed the *balanced content constraint*.

For the case when one is allowed to have arbitrary ℓ-substrings, the problem of enumerating profile vectors was independently addressed by Jacquet *et al.* [11] in the context of "Markov types". However, the method of Jacquet *et al.* does not extend to the case of enumeration of profiles with specific ℓ-substring constraints. To address this, we cast our more general enumeration question as a problem of *enumerating integer points in a rational polytope* and use *Ehrhart theory* to provide estimates of these values.

The paper is organized as follows. Section 2 introduces the notion of a sequence profile and the underlying balanced constraints and forbidden substring constraint. To enumerate constrained sequence profiles, we introduce the notion of a constrained de Bruijn graph in Section 3. Section 4 is devoted to the proof of the main enumeration results using Ehrhart theory. Numerical results for the balanced constraint enumeration problem are given in Section 5. Due to space constraints, certain proofs are omitted and the full proofs are found in the our preprint [12].

2. Profile Vectors

Let $[\![q]\!]$ denote the set of integers $\{0, 1, 2, \ldots, q-1\}$ and consider a word $\mathbf{x}$ of length n over $[\![q]\!]$. Suppose that $\ell < n$. An *ℓ-gram* is a substring of $\mathbf{x}$ of length ℓ.

Let $\mathbf{p}(\mathbf{x}; q, \ell)$ denote the (ℓ-gram) *profile vector*, i.e., a vector of length q^ℓ indexed by all vectors of $[\![q]\!]^\ell$ ordered lexicographically. In the profile vector, an entry indexed by $\mathbf{z}$ contains the number of occurrences of $\mathbf{z}$ as an ℓ-gram of $\mathbf{x}$. For example, $\mathbf{p}(0000; 2, 2) = (3, 0, 0, 0)$, while $\mathbf{p}(0101; 2, 2) = (0, 2, 1, 0)$. Observe that for any $\mathbf{x} \in [\![q]\!]^\ell$, the sum of entries in $\mathbf{p}(\mathbf{x}; q, \ell)$ equals $(n - \ell + 1)$.

Consider further a subset $S \subseteq [\![q]\!]^\ell$. For $\mathbf{x} \in [\![q]\!]^n$, we define the S-restricted profile $\mathbf{p}(\mathbf{x}; S)$ as a vector indexed by elements of S, whose entry corresponding to a sequence $\mathbf{z}$ gives the number of occurrences of $\mathbf{z}$ as an ℓ-gram of $\mathbf{x}$. Although arbitrary sequences may have substrings that lie outside S, for reasons to be apparent from our subsequent discussion, we focus only on vectors $\mathbf{x}$ whose ℓ-grams belong to S. As for the unrestricted case, the sum of the entries in $\mathbf{p}(\mathbf{x}; S)$ equals $n - \ell + 1$.

The choice of S is governed by certain reliability considerations in DNA storage sequence designs, including

This work was supported in part by the NSF STC Class 2010 CCF 0939370 grant and the Strategic Research Initiative (SRI) Grant conferred by the University of Illinois, Urbana-Champaign. Research of the second author was supported by the IC Postdoctoral Research Fellowship. This work was completed when the first author was at University of Illinois, Urbana Champaign.

[1]gBlocks for example is able to synthesize strings up to 2000bp [9].

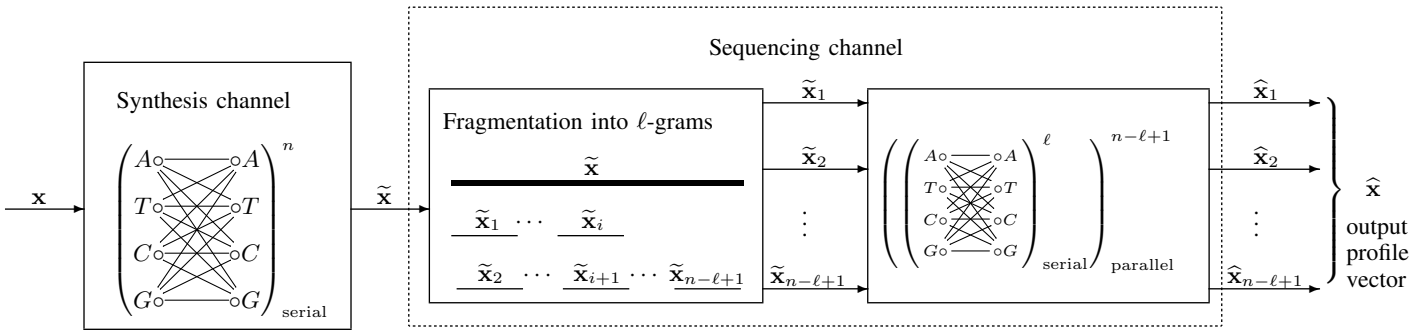


Fig. 1. The DNA Storage Channel. Information is encoded in a DNA sequence $\mathbf{x}$ which is synthesized with potential errors. The output of the synthesis process is $\widetilde{\mathbf{x}}$. During readout, the sequence $\widetilde{\mathbf{x}}$ is read through the sequencing channel, which fragments the sequence and possibly perturbs the fragments via substitution error. The output of the channel is a set of DNA fragments, along with their frequency count.

(i) **Weight profiles of ℓ-grams**. For the application at hand, one may want to choose S to consist of ℓ-grams with a fixed proportion of C and G bases, as this proportion – known as the GC-content of the sequence – influences the thermostability, folding processes and overall coverage of the ℓ-grams. From the perspective of sequencing, GC contents of roughly 50% are desired.
To make this modeling assumption more precise and general, we assume sets S of the form described below. Suppose that $0 \leq w_1 < w_2 \leq \ell$ and $1 \leq q_1 \leq q-1$. Let $[w_1, w_2]$ and $[q_1]$ denote the set of integers $\{w_1, w_1+1, \ldots, w_2\}$ and $\{1, 2, \ldots, q_1\}$, respectively. For each $\mathbf{x} \in [\![q]\!]^\ell$, let the $[q_1]$ *weight* of $\mathbf{x}$ be the number of symbols in $\mathbf{x}$ that belong to $[q_1]$, and denote it by $\mathrm{wt}(\mathbf{x}; q_1)$. Let

$$S(q, \ell; q_1, [w_1, w_2]) \triangleq \left\{ \mathbf{x} \in [\![q]\!]^\ell : \mathrm{wt}(\mathbf{x}; q_1) \in [w_1, w_2] \right\}$$

be the set of all sequences with ℓ-gram weights restricted to $[w_1, w_2]$. For example, by representing G, C, A, T by $1, 2, 3, 4$, respectively, by choosing $q = 4$ and $q_1 = 2$, the choice $w_1 = \lfloor \ell/2 \rfloor$, $w_2 = w_1 + 1$ ensures the balanced GC constraint. Also, note that $S(q, \ell; q-1, [0, \ell])$ equals $[\![q]\!]^\ell$.

(ii) **Forbidden ℓ-grams**. Studies have indicated that certain substrings in DNA sequences – such as GCG, CGC – are likely to cause sequencing errors (see [13]). Hence, one may also choose S so as to avoid certain ℓ-grams. Treatment of specialized sets of forbidden ℓ-grams is beyond the scope of this paper and is deferred to future work.

An appropriate choice of S may lower the probability of errors due to synthesis, lack of coverage and sequencing. For generality, we present all our subsequent results for the weight constraints of (i), rather than the special case of GC balanced sequences.

A. Problem Statement

Fix $S \subseteq [\![q]\!]^\ell$ and let $([\![q]\!]^n; S)$ denote all q-ary words of length n whose ℓ-grams belong to S. We define an equivalence relation on $([\![q]\!]^n; S)$ as follows: let $\mathbf{x} \sim \mathbf{y}$ if and only if $\mathbf{p}(\mathbf{x}; S) = \mathbf{p}(\mathbf{y}; S)$. Denote the set of equivalence classes under this relation by $\mathcal{Q}(n; S)$. We choose a *representative word* for each class and henceforth refer to elements in $\mathcal{Q}(n; S)$ through their representative words. Let $\mathbf{p}\mathcal{Q}(n; S)$ denote the set of profile vectors of words in $\mathcal{Q}(n; S)$, so that $|\mathbf{p}\mathcal{Q}(n; S)| = |\mathcal{Q}(n; S)|$.

Remark 1. In the case where $S = [\![q]\!]^\ell$, given a word $\mathbf{x}$, Ukkonen made certain observations on the words in the equivalence class of $\mathbf{x}$, but was unable to completely characterize all words in the class [14]. In this work, we focus on computing the *number* of equivalence classes for a general subset S.

3. De Bruijn Graphs and Enumeration Results

We use standard terminology from graph theory, following Bollobás [15]. A *directed graph (digraph)* D is a pair of sets (V, E), where V is the set of *nodes* and E is a set of ordered pairs of V, called *arcs*. If $e = (v, v')$ is an arc, we call v the *initial* node and v' the *terminal* node. We allows loops in our digraphs: in other words, we allow $v = v'$.

The *incidence matrix* of a digraph D is a matrix $\mathbf{B}(D)$ in $\{-1, 0, 1\}^{V \times E}$, where

$$\mathbf{B}(D)_{v,e} = \begin{cases} 1 & \text{if } e \text{ is not a loop and } v \text{ is its terminal node,} \\ -1 & \text{if } e \text{ is not a loop and } v \text{ is its source node,} \\ 0 & \text{otherwise.} \end{cases}$$

Observe that when a digraph D has loops, its incidence matrix $\mathbf{B}(D)$ has $\mathbf{0}$-columns indexed by these loops. When D is connected, then the rank of $\mathbf{B}(D)$ equals $|V| - 1$ [15].

A *walk* of length n in a digraph is a sequence of nodes $v_0 v_1 \cdots v_n$ such that $(v_i, v_{i+1}) \in E$ for all $i \in [\![n]\!]$. A walk is *closed* if $v_0 = v_n$ and a *cycle* is a closed walk with distinct arcs and nodes, i.e. $(v_i, v_{i+1}) \neq (v_j, v_{j+1})$ and $v_i \neq v_j$, for $0 \leq i < j \leq n$. A loop corresponds to a cycle of length one. A closed walk is *Eulerian* if it includes all arcs in E.

Given a subset C of the arc set, let $\boldsymbol{\chi}(C) \in \{0,1\}^E$ be its *incidence vector*, where $\boldsymbol{\chi}(C)_e$ is one if $e \in C$ and zero otherwise. In general, for any closed walk C in D, we have $\mathbf{B}(D)\boldsymbol{\chi}(C) = \mathbf{0}$. Note that if for all $\mathbf{z}, \mathbf{z}' \in V(S)$, there exists a directed path from $\mathbf{z}$ to $\mathbf{z}'$ and vice versa, the digraph is termed *strongly connected*.

We are concerned with a generalization of the well known de Bruijn graphs [16]. Given q and ℓ, the standard *de Bruijn graph* is defined on the set $[\![q]\!]^\ell$. Here, we instead define the graph on a subset $S \subseteq [\![q]\!]^\ell$ and refer to the corresponding graph as *a restricted de Bruijn graph*, denoted by $D(S)$. The nodes of $D(S)$ are the $(\ell-1)$-prefixes and -suffixes of words in S. The pair $(\mathbf{v}, \mathbf{v}')$ belongs to the arc set if and only if $v_i = v'_{i-1}$ for $2 \leq i \leq \ell$ and $v_1 v_2 \cdots v_{\ell-1} v'_{\ell-1} \in S$. We identify the arc set with S and denote the node set as $V(S)$.

The notion of restricted de Bruijn graphs was introduced by Ruskey *et al.* [17] for the case of a binary alphabet. In their work, Ruskey *et al.* showed that $D(S)$ is Eulerian when $S =$

$S(2, \ell; 1, [w-1, w])$, for $w \in [\ell]$. Their results may be extended to general values of q, q_1, and more general range of weights. For simplicity, we abbreviate $D(S(q, \ell; q_1, [w_1, w_2]))$ and $D(\llbracket q \rrbracket^\ell)$ to $D(q, \ell; q_1, [w_1, w_2])$ and $D(q, \ell)$, respectively.

Proposition 3.1. Fix q and ℓ. Let $1 \le q_1 \le q-1$ and $1 \le w_1 < w_2 \le \ell$. Then $D(q, \ell; q_1, [w_1, w_2])$ is Eulerian.

Observe that when $q_1 = q-1$, $w_1 = 0$, $w_2 = \ell$, we recover the classical result that the de Bruijn graph $D(q, \ell)$ is Eulerian and Hamiltonian.

A. Enumerating $\mathcal{Q}(n; S)$

We provide the main enumeration results for $\mathcal{Q}(n; S)$, or equivalently, for $\mathbf{p}\mathcal{Q}(n; S)$. We first assume that $D(S)$ is strongly connected. In addition, we consider closed walks in $D(S)$, or equivalently, *closed words* that start and end with the same $(\ell-1)$-gram. We denote the set of closed words in $\mathcal{Q}(n; S)$ by $\bar{\mathcal{Q}}(n; S)$, and the corresponding set of profile vectors by $\mathbf{p}\bar{\mathcal{Q}}(n; S)$.

Suppose that $\mathbf{u}$ belongs to $\mathbf{p}\bar{\mathcal{Q}}(n; S)$. Then the following system of linear equations that we refer to as the *flow conservation equations*, hold true: $\mathbf{B}(D(S))\mathbf{u} = \mathbf{0}$. Let $\mathbf{1}$ denote the all-ones vector. Since the number of ℓ-grams in a word of length n is $n-\ell+1$, we also have $\mathbf{1}^T\mathbf{u} = n-\ell+1$.

Let $\mathbf{A}(S)$ be $\mathbf{B}(D(S))$ augmented with a top row $\mathbf{1}^T$; let $\mathbf{b}$ be a vector of length $|V(S)|+1$ with a one as its first entry, and zeros elsewhere. Then the constraint equations may be written as

$$\mathbf{A}(S)\mathbf{u} = (n-\ell+1)\mathbf{b}.$$

Consider the following two sets of integer points,

$$\mathcal{F}(n; S) \triangleq \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}(S)\mathbf{u} = (n-\ell+1)\mathbf{b},\ \mathbf{u} \ge \mathbf{0}\}, \quad (1)$$

$$\mathcal{E}(n; S) \triangleq \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}(S)\mathbf{u} = (n-\ell+1)\mathbf{b},\ \mathbf{u} > \mathbf{0}\}. \quad (2)$$

The preceding discussion asserts that any profile vector must lie in $\mathcal{F}(n; S)$. Conversely, the next lemma shows that any vector in $\mathcal{E}(n; S)$ is a profile vector of some word in $\bar{\mathcal{Q}}(n; S)$.

Lemma 3.2. Suppose that $D(S)$ is strongly connected. If $\mathbf{u} \in \mathcal{E}(n; S)$, then there exists a word $\mathbf{x} \in \bar{\mathcal{Q}}(n; S)$ such that $\mathbf{p}(\mathbf{x}; S) = \mathbf{u}$ or $\mathbf{u} \in \mathbf{p}\bar{\mathcal{Q}}(n; S)$.

Therefore, we have the following relation,

$$\mathcal{E}(n; S) \subseteq \mathbf{p}\bar{\mathcal{Q}}(n; S) \subseteq \mathcal{F}(n; S). \quad (3)$$

We first state our main enumeration result and defer its proof to Section 4. Specifically, under the assumption that $D(S)$ is strongly connected, we show that both $|\mathcal{E}(n; S)|$ and $|\mathcal{F}(n; S)|$ are quasipolynomials in n whose coefficients are periodic in n. Formally, we define a *quasipolynomial* f as a function in t of the form $c_D(t)t^D + c_{D-1}(t)t^{D-1} + \cdots + c_0(t)$, where $c_D, c_{D-1}, \ldots, c_0$ are periodic functions of t. If c_D is not identically equal to zero, f is said to be of *degree* D. The *period* of f is given by the lowest common multiple of the periods of $c_D, c_{D-1}, \ldots, c_0$.

In the theorem, we used standard asymptotic notation. However, we adapt the Ω and Θ symbols in order to succinctly present our results. We use $f(n) = \Omega'(g(n))$ to state that for a fixed value of ℓ, there exists an integer λ and a positive constant c so that $f(n) \ge cg(n)$ for sufficiently large n with $\lambda | (n-\ell+1)$. Furthermore, $f(n) = \Theta'(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega'(g(n))$.

Theorem 3.3. Suppose $D(S)$ is strongly connected and let λ be the lowest common multiple of the lengths of all cycles in $D(S)$. Then $|\mathcal{E}(n; S)|$ and $|\mathcal{F}(n; S)|$ are both quasipolynomials in n of the same degree $|S| - |V(S)|$ and share the same period that divides λ. In particular, $|\mathbf{p}\bar{\mathcal{Q}}(n; S)| = \Theta'\left(n^{|S|-|V(S)|}\right)$.

Before we end this section, we look at certain implications of Theorem 3.3. We also provide estimates for $|\mathbf{p}\mathcal{Q}(n; S)|$ when $D(S)$ is strongly connected, and for $|\mathbf{p}\mathcal{Q}(n; S)|$ and $|\mathbf{p}\bar{\mathcal{Q}}(n; S)|$, when $D(S)$ is not necessarily strongly connected.

Corollary 3.4. Suppose $D(S)$ is strongly connected. For any $\mathbf{z}, \mathbf{z}' \in V(S)$, consider the set of words in $\mathcal{Q}(n; S)$ that begin with $\mathbf{z}$ and end with $\mathbf{z}'$, and let $\mathbf{p}\mathcal{Q}(n; S, \mathbf{z} \to \mathbf{z}')$ be the corresponding set of profile vectors. Similarly, let $\mathbf{p}\mathcal{Q}(n; S, \mathbf{z} \to *)$ and $\mathbf{p}\mathcal{Q}(n; S, * \to \mathbf{z}')$ denote the set of profile vectors of words beginning with $\mathbf{z}$ and words ending with $\mathbf{z}'$, respectively. Then

$$\begin{aligned}|\mathbf{p}\mathcal{Q}(n; S)| &= \Theta'(|\mathbf{p}\mathcal{Q}(n; S, \mathbf{z} \to \mathbf{z}')|) = \Theta'(|\mathbf{p}\mathcal{Q}(n; S, * \to \mathbf{z}')|) \\ &= \Theta'(|\mathbf{p}\mathcal{Q}(n; S, \mathbf{z} \to *)|) = \Theta'\left(n^{|S|-|V(S)|}\right).\end{aligned}$$

In the special case where $S = \llbracket q \rrbracket^\ell$, Jacquet *et al.* demonstrated a stronger version of Theorem 3.3 using analytic combinatorics. In addition, using a careful analysis similar to the proof of Corollary 3.4, Jacquet *et al.* also provided a tighter bound for $|\mathbf{p}\mathcal{Q}(n; q, \ell)|$ for one choice of the parameters, $\ell = 2$. Note that $f(n) \sim g(n)$ stands for $\lim_{n\to\infty} f(n)/g(n) = 1$.

Theorem 3.5 (Jacquet *et al.* [11])**.** Fix q, ℓ. Let $\mathcal{E}(n; \llbracket q \rrbracket^\ell)$, $\mathcal{F}(n; \llbracket q \rrbracket^\ell)$, $\mathbf{p}\mathcal{Q}(n; q, \ell)$ and $\mathbf{p}\bar{\mathcal{Q}}(n; q, \ell)$ be defined as above. Then

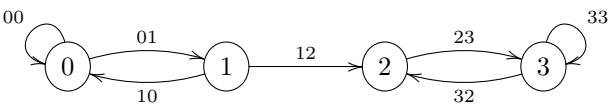
$$|\mathcal{E}(n; \llbracket q \rrbracket^\ell)| \sim |\mathcal{F}(n; \llbracket q \rrbracket^\ell)| \sim |\mathbf{p}\bar{\mathcal{Q}}(n, q, \ell)| \sim c(q, \ell) n^{q^\ell - q^{\ell-1}}, \quad (4)$$

where $c(q, \ell)$ is a constant. Furthermore, when $\ell = 2$, we have $|\mathbf{p}\mathcal{Q}(n; q, 2)| = (q^2 - q + 1)|\mathbf{p}\bar{\mathcal{Q}}(n; q, 2)|(1 - O(n^{-2q}))$.

Next, we extend Theorem 3.3 to provide estimates on $\bar{\mathcal{Q}}(n; S)$ and $\mathcal{Q}(n; S)$ for general digraphs.

Corollary 3.6. Given $D(S)$, let $V_1, V_2, \ldots, V_I$ be a partition of $V(S)$, such that the induced subgraph (V_i, S_i) is a maximal strongly connected component for all $i \in I$. Define $\bar{\Delta} \triangleq \max\{|S_i| - |V_i| : i \in I\}$. Then $|\bar{\mathcal{Q}}(n; S)| = \Theta'(n^{\bar{\Delta}})$.

Example 3.1. Let $S = \{00, 01, 10, 12, 23, 32, 33\}$ with $q = 4$ and $\ell = 2$. Then $D(S)$ is as shown below.



We have two strongly connected components, namely, $V_1 = \{0, 1\}$ and $V_2 = \{2, 3\}$. So, $(V_1, S_1 = \{00, 01, 10\})$ and $(V_2, S_2 = \{23, 32, 33\})$ are both strongly connected digraphs with $|\mathbf{p}\bar{\mathcal{Q}}(n; S_1)| = |\mathbf{p}\bar{\mathcal{Q}}(n; S_2)| = \lceil n/2 \rceil = \Theta'(n)$. Hence, $|\mathbf{p}\bar{\mathcal{Q}}(n; S)| = |\mathbf{p}\bar{\mathcal{Q}}(n; S_1)| + |\mathbf{p}\bar{\mathcal{Q}}(n; S_2)| = \Theta'(n)$, in agreement with Corollary 3.6.

On the other hand, let us enumerate the elements of $\mathcal{Q}(n; S)$ or $\mathbf{p}\mathcal{Q}(n; S)$. Let $\mathbf{u} \in \mathbf{p}\mathcal{Q}(n; S)$. If $u_{12} = 0$, then $\mathbf{u}$ belongs to $\mathbf{p}\mathcal{Q}(n; S_1)$ or $\mathbf{p}\mathcal{Q}(n; S_2)$. Otherwise, $u_{12} = 1$ and we have $\mathbf{u} = \mathbf{u}_1 + \boldsymbol{\chi}(12) + \mathbf{u}_2$ with $\mathbf{u}_1 \in \mathbf{p}\mathcal{Q}(n_1 + 1; S_1, * \to 1)$, $\mathbf{u}_2 \in \mathbf{p}\mathcal{Q}(n_2 + 1; S_2, 2 \to *)$ and $n_1 + n_2 + 1 = n - 1$. Now, $|\mathbf{p}\mathcal{Q}(n; S_1)| = |\mathbf{p}\mathcal{Q}(n; S_2)| = n + \lfloor n/2 \rfloor$ and $|\mathbf{p}\bar{\mathcal{Q}}(n; S_1, * \to$

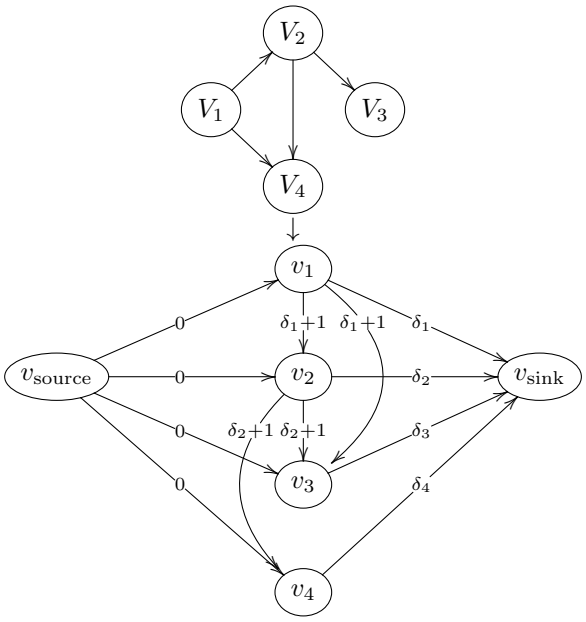


Fig. 2. Constructing a weighted digraph from the strongly connected components of $D(S)$.

$1)| = |\mathbf{p}\bar{\mathcal{Q}}(n; S_2, 2 \to *)| = n - 1$ for $n \geq 2$. Hence,

$$|\mathbf{p}\mathcal{Q}(n;S)| = 2\left(n + \left\lfloor \frac{n}{2} \right\rfloor\right) + 2(n-2) + \sum_{n_1=1}^{n-3} n_1(n-2-n_1) = \Theta'(n^3).$$

Therefore, when $D(S)$ is not strongly connected, it is not necessarily true that $|\mathbf{p}\bar{\mathcal{Q}}(n;S)|$ and $|\mathbf{p}\mathcal{Q}(n;S)|$ differ only by a constant factor. Furthermore, we can extend the methods in this example to obtain $|\mathbf{p}\mathcal{Q}(n;S)|$ for general digraphs.

To determine $|\mathbf{p}\mathcal{Q}(n;S)|$, we construct an auxiliary weighted digraph with nodes $v_1, v_2, \ldots, v_I, v_{\text{source}}$ and v_{sink}. If there exists an arc from the component V_i to component V_j, $i, j \in [I]$, we add an arc from v_i to v_j. Furthermore, we add an arc from v_{source} to v_i and from v_i to v_{sink} for all $i \in [I]$. The arcs leaving v_{source} have zero weight. For all $i \in [I]$, the arcs leaving v_i have weight $\delta_i = |S_i| - |V_i|$ if their terminal node is v_{sink}, and weight $\delta_i + 1$ otherwise (see Fig. 2 for the transformation).

Let D' be the resulting digraph and observe that D' is acyclic. Hence, we can find the longest weighted path from v_{source} to v_{sink} in linear time. Suppose that Δ is the weight of the longest path. Then the next corollary states that $|\mathbf{p}\mathcal{Q}(n;S)| = \Theta'(n^{\Delta})$.

Corollary 3.7. Given $D(S)$, let $V_1, V_2, \ldots, V_I$ be a partition of $V(S)$ such that the induced subgraphs (V_i, S_i) are strongly connected for all $i \in I$. Construct D' as above (see Fig. 2) and let Δ be the weight of the longest weighted path from v_{source} to v_{sink}. Then, $|\mathbf{p}\mathcal{Q}(n;S)| = \Theta'(n^{\Delta})$.

4. Ehrhart Theory and Proof of Theorem 3.3

We assume $D(S)$ to be strongly connected and provide a detailed proof of Theorem 3.3. For this purpose, we introduce some fundamental results from Ehrhart theory. Ehrhart theory is a natural framework for enumerating profile vectors and one may simplify the proof of [11] significantly and generalize the corresponding results to a bigger family of digraphs. Furthermore, Ehrhart theory also allows us to extend the enumeration procedure to profiles at a prescribed distance (see [10]).

As hinted by (1) and (2), to enumerate codewords of interest, we need to enumerate certain sets of lattice points in polytopes. The first general treatment of the theory of enumerating lattice points in polytopes was described by Ehrhart [18]. Here, we follow the combinatorial treatment of Beck and Robins [19].

Consider any *rational polytope* $\mathcal{P}$ given by $\mathcal{P} \triangleq \{\mathbf{u} \in \mathbb{R}^n : \mathbf{A}\mathbf{u} \leq \mathbf{b}\}$, for some integer matrix $\mathbf{A}$ and some integer vector $\mathbf{b}$. A rational polytope is *integer* if all its vertices are integral. The *lattice point enumerator* $L_{\mathcal{P}}(t)$ of $\mathcal{P}$ is given by $L_{\mathcal{P}}(t) \triangleq \#(\mathbb{Z}^n \cap t\mathcal{P})$ for all $t \in \mathbb{Z}_{>0}$.

Ehrhart [18] introduced the lattice point enumerator for rational polytopes and showed that $L_{\mathcal{P}}(t)$ is a quasipolynomial of degree D, where D is given by the dimension of the polytope $\mathcal{P}$. Here, we define the *dimension* of a polytope to be the dimension of the affine space spanned by points in $\mathcal{P}$. A formal statement of Ehrhart's theorem is provided below.

Theorem 4.1 (Ehrhart's theorem for polytopes [19, Thm 3.8 and 3.23])**.** If $\mathcal{P}$ is a rational convex polytope of dimension D, then $L_{\mathcal{P}}(t)$ is a quasipolynomial of degree D. Its period divides the least common multiple of the denominators of the coordinates of the vertices of $\mathcal{P}$. Furthermore, if $\mathcal{P}$ is integer, then $L_{\mathcal{P}}(t)$ is a polynomial of degree D.

Motivated by (2), we consider the *relative interior* of $\mathcal{P}$. For the case when $\mathcal{P}$ is convex, the relative interior, or interior, is given by $\mathcal{P}^{\circ} \triangleq \{\mathbf{u} \in \mathcal{P} : \text{for all } \mathbf{u}' \in \mathcal{P}, \text{there exists an } \epsilon > 0 \text{ such that } \mathbf{u} + \epsilon(\mathbf{u} - \mathbf{u}') \in \mathcal{P}\}$.

For a positive integer t, we consider the quantity $L_{\mathcal{P}^{\circ}}(t) = \#(\mathbb{Z}^n \cap t\mathcal{P}^{\circ})$. Ehrhart conjectured the following relation between $L_{\mathcal{P}}(t)$ and $L_{\mathcal{P}^{\circ}}(t)$, which was proved by Macdonald [20].

Theorem 4.2 (Ehrhart-Macdonald reciprocity [19, Thm 4.1])**.** If $\mathcal{P}$ is a rational convex polytope of dimension D, then the evaluation of $L_{\mathcal{P}}(t)$ at negative integers satisfies $L_{\mathcal{P}}(-t) = (-1)^D L_{\mathcal{P}^{\circ}}(t)$.

Recall the definitions of $\mathbf{A}(S)$ and $\mathbf{b}$ in (1), and consider the polytope

$$\mathcal{P}(S) \triangleq \{\mathbf{u} \in \mathbb{R}^{|S|} : \mathbf{A}(S)\mathbf{u} = \mathbf{b}, \mathbf{u} \geq \mathbf{0}\}. \tag{5}$$

Using lattice point enumerators, we may write $|\mathcal{F}(n;S)| = L_{\mathcal{P}(S)}(n - \ell + 1)$. Therefore, in view of Ehrhart's theorem, we determine the dimension of the polytope $\mathcal{P}(S)$ and characterize its interior and its vertices in the following technical lemma.

Lemma 4.3. Suppose that $D(S)$ is strongly connected. Then the following properties of $\mathcal{P}(S)$ hold.

(P1) The dimension of $\mathcal{P}(S)$ is $|S| - |V(S)|$.

(P2) $\mathcal{P}^{\circ}(S) = \{\mathbf{u} \in \mathbb{R}^{|S|} : \mathbf{A}(S)\mathbf{u} = \mathbf{b}, \mathbf{u} > \mathbf{0}\}$ and therefore, $|\mathcal{E}(n;S)| = L_{\mathcal{P}^{\circ}(S)}(n - \ell + 1)$.

(P3) The vertex set of $\mathcal{P}(S)$ is given by $\{\boldsymbol{\chi}(C)/|C| : C \text{ is a cycle in } D(S)\}$.

Therefore, using Ehrhart's theorem and Ehrhart-Macdonald reciprocity along with (P1) and (P2), we arrive at the fact that $|\mathcal{E}(n;S)|$ and $|\mathcal{F}(n;S)|$ are quasipolynomials in n whose coefficients are periodic in n.

Let λ_S be the lowest common multiple of the lengths of all cycles in $D(S)$. Then the period of the quasipolynomial $L_{\mathcal{P}(S)}(n - \ell + 1)$ divides λ_S by Ehrhart's theorem and (P3).

Let us dilate the polytope $\mathcal{P}(S)$ by λ_S and consider the enumerator $L_{\lambda_S\mathcal{P}(S)}(t)$. Since $\lambda_S\mathcal{P}$ is integer, both $L_{\lambda_S\mathcal{P}(S)}(t)$ and $L_{\lambda_S\mathcal{P}^{\circ}(S)}(t)$ are polynomials of degree $|S| - |V(S)|$. Hence,

whenever $n-\ell+1 = \lambda_S t$ or $\lambda_S|(n-\ell+1)$, $|\bar{\mathcal{Q}}(n;S)| \geq L_{\lambda_S \mathcal{P}^\circ(S)}(t) = \Omega\left(t^{|S|-|V(S)|}\right)$, and therefore, $|\bar{\mathcal{Q}}(n;S)| = \Theta'\left(n^{|S|-|V(S)|}\right)$. This completes the proof of Theorem 3.3.

Finally, when $D(S)$ contains loops, we can further show that the leading coefficients of the quasipolynomials $|\mathcal{E}(n;[\![q]\!]^\ell)|$ and $|\mathcal{F}(n;[\![q]\!]^\ell)|$ are the same and constant (i.e. aperiodic). This result is a direct consequence of Ehrhart-Macdonald reciprocity and the fact that $|\mathcal{E}(n;[\![q]\!]^\ell)|$ is monotonically increasing. When $S = [\![q]\!]^\ell$, Corollary 4.4 yields (4), a result of Jacquet *et al.* [11].

Corollary 4.4. Suppose $D(S)$ is strongly connected. If $D(S)$ has loops, then for some constant $c(S)$.

$$|\mathcal{E}(n;S)| \sim |\bar{\mathcal{Q}}(n;S)| \sim |\mathcal{F}(n;S)| \sim c(S) n^{|S|-|V(S)|} + O(n^{|S|-|V(S)|-1}).$$

5. Numerical Computations for $S(q,\ell;q_1,[w1,w2])$

We summarize numerical results for code sizes pertaining to the special case when $S = S(q,\ell;q_1,[w_1,w_2])$.

By Proposition 3.1, $D(q,\ell;q_1,[w_1,w_2])$ is Eulerian and therefore strongly connected. In other words, Theorem 3.3 applies and we have $|\mathcal{Q}(n;S)| = \Theta'(n^{|S|-|V(S)|})$, where $|S| = \sum_{w=w_1}^{w_2} \binom{\ell}{w} q_1^w (q-q_1)^{\ell-w}$, while $|V(S)|$ is given by $|S(q,\ell-1;q_1,[w_1-1,w_2])| = \sum_{w=w_1-1}^{w_2} \binom{\ell-1}{w} q_1^w (q-q_1)^{\ell-1-w}$.

Let $D = |S| - |V(S)|$. We determine next the coefficient of n^D in $|\mathcal{Q}(n;S)|$. When $w_2 = \ell$, the digraph $D(q,\ell;q_1,[w_1,\ell])$ contains the loop that correspond to the ℓ-gram profile $\mathbf{1}^T$. Hence, by Corollary 4.4, the desired coefficient is aperiodic and we denote it by $c(q,\ell;q_1,[w_1,\ell])$. When $S = [\![q]\!]^\ell$, we denote this coefficient by $c(q,\ell)$ and remark that this value corresponds to the constant defined in Theorem 3.5.

When $w_2 < \ell$, the digraph $D(q,\ell;q_1,[w_1,w_2])$ does not contain any loops. Recall from Section 4 the definitions of $\mathcal{P}(S)$, λ_S and $L_{\mathcal{P}(S)}(n-\ell+1)$. In particular, recall that the lattice point enumerator $L_{\mathcal{P}(S)}(n-\ell+1)$ is a quasipolynomial of degree D whose period divides λ_S and that consequently, the coefficient of n^D in $|\mathcal{Q}(n;S)|$ is periodic. For ease of presentation, we only determine the coefficient of n^D for those parameter values for which λ_S divides $(n-\ell+1)$, i.e., for which $n-\ell+1 = \lambda_S t$, for some integer t. Then, the desired coefficient is given by $c(q,\ell;q_1,[w_1,w_2]) \triangleq c/\lambda_S^D$, where c is the leading coefficient of the polynomial $L_{\lambda_S \mathcal{P}(S)}(t)$. Hence, we have the following result.

Corollary 5.1. Consider $S = S(q,\ell;q_1,[w_1,w_2])$ and define $D = \sum_{w=w_1}^{w_2} \binom{\ell}{w} q_1^w (q-q_1)^{\ell-w} - \sum_{w=w_1-1}^{w_2} \binom{\ell-1}{w} q_1^w (q-q_1)^{\ell-1-w}$. Suppose that $\lambda_S = \mathrm{lcm}\{|C| : C \text{ is a cycle in } D(S)\}$. Then for some constant $c(q,\ell;q_1,[w_1,w_2])$,

(i) if $w_2 = \ell$, $|\mathcal{Q}(n;S)| = c(q,\ell;q_1,[w_1,\ell]) n^D + O(n^{D-1})$ for all n;

(ii) otherwise, $|\mathcal{Q}(n;S)| = c(q,\ell;q_1,[w_1,w_2]) n^D + O(n^{D-1})$ for all n such that $\lambda_S|(n-\ell+1)$.

We determine $c(q,\ell;q_1,[w_1,w_2])$ via numerical computations. Computing the lattice point enumerator is a fundamental problem in discrete optimization and many algorithms and software implementations have been developed for such purposes. We make use of the software `LattE`, developed by Baldoni *et al.* [21], which is based on an algorithm of Barvinok [22].

Using `LattE`, we computed the desired coefficients for various values of $(q,\ell;q_1,[w_1,w_2])$. As an illustrative example, `LattE` determined $c(2,4) = 283/9754214400$ with computational time less than a minute. This shows that although the exact evaluation of $c(q,\ell)$ is prohibitively complex (as pointed by Jacquet *et al.* [11]), numerical computations of $c(q,\ell)$ and $c(q,\ell;q_1,[w_1,w_2])$ are feasible for certain moderate values of parameters. We tabulate $c(q,\ell;q_1,[w_1,w_2])$ in Table I.

TABLE I
Computation of $c(q,\ell;q_1,[w_1,w_2])$. We fixed $q=2$ and $q_1=1$.

ℓ	w_1	w_2	D	λ_S	$c(2,\ell;1,[w_1,w_2])$
4	2	3	3	60	1/360
4	2	4	4	–	1/1440
5	2	3	6	120	1/5184000
5	2	4	10	27720	40337/34566497280000000
5	2	5	11	–	3667/34566497280000000
5	3	4	4	420	23/302400
5	3	5	5	–	23/1512000
6	3	4	10	65520	43919/754932300595200000
6	4	5	5	840	1/518400

References

[1] S. Kannan and A. McGregor, "More on reconstructing strings from random traces: insertions and deletions," in *Proc. IEEE Intl. Inform. Theory*. IEEE, 2005, pp. 297–301.

[2] J. Acharya, H. Das, O. Milenkovic, A. Orlitsky, and S. Pan, "On reconstructing a string from its substring compositions," in *Proc. IEEE Intl. Symp. Inform. Theory*. IEEE, 2010, pp. 1238–1242.

[3] ——, "Quadratic-backtracking algorithm for string reconstruction from substring compositions," in *Proc. IEEE Intl. Symp. Inform. Theory*. IEEE, 2014, pp. 1296–1300.

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

[5] 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*, 2013.

[6] J. Ma, O. Milenkovic, and H. Zhao, "Strategic Research Initiative (SRI) grant, Rewritable DNA storage," Patent, pending.

[7] P. Medvedev, K. Georgiou, G. Myers, and M. Brudno, "Computability of models for sequence assembly," in *Algorithms in Bioinformatics*. Springer, 2007, pp. 289–301.

[8] P. E. Compeau, P. A. Pevzner, and G. Tesler, "How to apply de Bruijn graphs to genome assembly," *Nature biotechnology*, vol. 29, no. 11, pp. 987–991, 2011.

[9] gBlocks Gene Fragments. [Online]. Available: http://www.idtdna.com/pages/products/genes/gblocks-gene-fragments

[10] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Coding for DNA storage channels," in *IEEE Inform. Theory Workshop*, 2015, to appear.

[11] 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.

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

[13] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara, and S. Kanaya, "Sequence-specific error profile of Illumina sequencers," *Nucleic acids research*, p. gkr344, 2011.

[14] E. Ukkonen, "Approximate string-matching with q-grams and maximal matches," *Theoretical computer science*, vol. 92, no. 1, pp. 191–211, 1992.

[15] B. Bollobás, *Modern graph theory*. Springer, 1998, vol. 184.

[16] N. G. de Bruijn, "A combinatorial problem," *Koninklijke Nederlandse Akademie v. Wetenschappen*, vol. 49, no. 49, pp. 758–764, 1946.

[17] F. Ruskey, J. Sawada, and A. Williams, "De Bruijn sequences for fixed-weight binary strings," *SIAM Journal on Discrete Mathematics*, vol. 26, no. 2, pp. 605–617, 2012.

[18] E. Ehrhart, "Sur les polyédres rationnels homothétiques á n dimensions," *CR Acad. Sci. Paris*, vol. 254, pp. 616–618, 1962.

[19] M. Beck and S. Robins, *Computing the continuous discretely: Integer-point enumeration in polyhedra*. Springer, 2007.

[20] I. G. Macdonald, "Polynomials associated with finite cell-complexes," *J. London Math. Society*, vol. 2, no. 1, pp. 181–192, 1971.

[21] V. Baldoni, N. Berline, J. A. De Loera, B. Dutra, M. Köppe, S. Moreinis, G. Pinto, M. Vergne, and J. Wu, "A user's guide for latte integrale v1. 7.1," *Optimization*, vol. 22, p. 2, 2014.

[22] A. I. Barvinok, "A polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed," *Mathematics of Operations Research*, vol. 19, no. 4, pp. 769–779, 1994.