Codes for DNA Storage Channels

Han Mao Kiah, Gregory J. Puleo, and Olgica Milenkovic
Coordinated Science Laboratory, University of Illinois, Urbana-Champaign

***Abstract*—We consider the problem of assembling a sequence based on a collection of its substrings observed through a noisy channel. This problem of reconstructing sequences from traces was first investigated in the noiseless setting under the name of "Markov type" analysis. Here, we explain the connection between the problem and the problem of DNA synthesis and sequencing, and introduce the notion of a DNA storage channel. We analyze the number of sequence equivalence classes under the channel mapping and propose new asymmetric coding techniques to combat the effects of synthesis noise. In our analysis, we make use of Ehrhart theory for rational polytopes.**

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 high capacity and low-maintanence 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 the setting we propose to analyze, one first synthesizes a sequence $\mathbf{x} \in \mathcal{D} = \{A, T, G, C\}^n$, and then fragments it in the process of sequencing into a collection of substrings of approximately the same length, ℓ. The length ℓ ranges anywhere between 100 to 1500 nts[1]. Ideally, one would like to synthesize $\mathbf{x}$ and sequence all ℓ-substrings without errors, which is not possible in practice. For large n, the synthesis error-rate of $\mathbf{x}$ is roughly $1-3\%$. Substrings of short length may be sequenced with an error-rate not exceeding 1%; long substrings exhibit much higher sequencing error-rates, often as high as 15%. Furthermore, due to non-uniform fragmentation, a number of the substrings are not available for sequencing, leaving coverage gaps in the original sequence.

To model this read-write phenomena, we introduce the notion of a *DNA storage channel* that takes as its input a sequence $\mathbf{x}$ of length n, introduces d_{syn} substitution errors in $\mathbf{x}$, resulting in a sequence $\tilde{\mathbf{x}}$. The channel proceeds to produce all or a subset of substrings of the sequence $\tilde{\mathbf{x}}$ of length ℓ, $\ell < n$. Each of the substrings is allowed to have additional substitution errors, due to sequencing. The number of substring sequencing errors equals d_{seq}[2]. The substrings at the output of the DNA storage channel are collectively enumerated by a vector $\hat{\mathbf{x}}$, termed the channel output (see Fig. 1 for an illustration).

The main contributions of the paper are as follows. The first contribution is to *model the read process (sequencing)* through the use of *profile vectors*. A profile vector of a sequence enumerates all substrings of the sequence, and profiles form a pseudometric space amenable for analysis. The second contribution of the paper is to *introduce a new family of codes* for two classes of errors arising in the DNA storage channel due to synthesis and lack of coverage, and show that they may be characterized by *asymmetric errors* studied in classical coding theory. Our third contribution is a code design technique which makes use of codewords with different profile vectors or profile vectors at sufficiently large distance from each other. The analysis in the former case amounts to enumerating all valid profile vectors, and this problem was independently addressed by Jacquet *et al.* [9] in the context of "Markov types". The method of Jacquet *et al.* does not extend to the case of enumeration of profiles at sufficiently large distance from each other. Instead, we cast the more general code design question as a problem of *enumerating integer points in a rational polytope* and use tools from *Ehrhart theory* to provide estimates of the sizes of the underlying codes.

2. Profile Vectors and a Metric Space

Let $[\![q]\!]$ denote the set of integers $\{0, 1, 2, \ldots, q-1\}$. Consider a vector or 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* of length $[\![q]\!]^\ell$, indexed by all vectors of $[\![q]\!]^\ell$ ordered lexicographically. In the profile vector, an entry indexed by $\mathbf{z}$ gives 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 $(n - \ell + 1)$.

Let N be an integer. Define the *L_1-weight* of a word $\mathbf{u} \in \mathbb{Z}_{\geq 0}^N$ as $\mathrm{wt}(\mathbf{u}) \triangleq \sum_{i=1}^N u_i$. In addition, for any pair of words $\mathbf{u}, \mathbf{v} \in \mathbb{Z}_{\geq 0}^N$, let $N(\mathbf{u}, \mathbf{v}) \triangleq \sum_{i=1}^N \max(u_i - v_i, 0)$ and define the *asymmetric distance* as $d_{\mathrm{asym}}(\mathbf{u}, \mathbf{v}) = \max(N(\mathbf{u}, \mathbf{v}), N(\mathbf{v}, \mathbf{u}))$. A set $\mathcal{C}$ is called an *(N, d)-asymmetric error correcting code (AECC)* if $\mathcal{C} \subseteq \mathbb{Z}_{\geq 0}^N$ and $d = \min\{d_{\mathrm{asym}}(\mathbf{x}, \mathbf{y}) : \mathbf{x}, \mathbf{y} \in \mathcal{C}, \mathbf{x} \neq \mathbf{y}\}$. For any $\mathbf{x} \in \mathcal{C}$, let $\mathbf{e} \in \mathbb{Z}_{\geq 0}^N$ be such that $\mathbf{x} - \mathbf{e} \geq \mathbf{0}$. We say that an *asymmetric error* $\mathbf{e}$ occurred if the received word is $\mathbf{x} - \mathbf{e}$. We have the following theorem characterizing asymmetric error-correction codes (see [10, Thm 9.1]).

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.

[1] For our system currently under development, due to the high cost of synthesis, we have chosen $n = 1000$. In addition, we used multiple sequences to increase storage capacity.

[2] In what follows, we limit our attention to the case $d_{seq} = 0$.

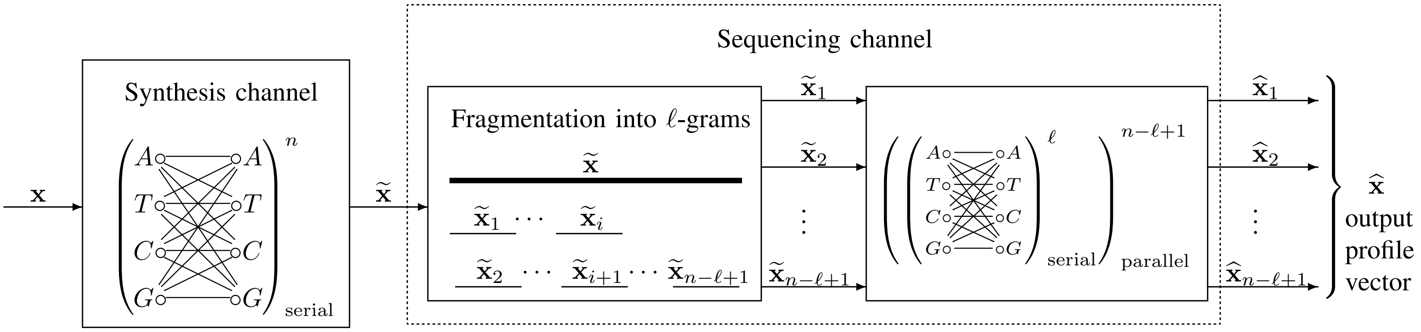
Sequencing channel
Synthesis channel
$\mathbf{x}$
$\widetilde{\mathbf{x}}$
n
serial
Fragmentation into ℓ-grams
$\widetilde{\mathbf{x}}$
$\widetilde{\mathbf{x}}_1 \cdots \widetilde{\mathbf{x}}_i$
$\widetilde{\mathbf{x}}_2 \cdots \widetilde{\mathbf{x}}_{i+1} \cdots \widetilde{\mathbf{x}}_{n-\ell+1}$
$\widetilde{\mathbf{x}}_1$
$\widetilde{\mathbf{x}}_2$
$\widetilde{\mathbf{x}}_{n-\ell+1}$
ℓ
$n-\ell+1$
serial
parallel
A T C G
$\widehat{\mathbf{x}}_1$
$\widehat{\mathbf{x}}_2$
$\widehat{\mathbf{x}}_{n-\ell+1}$
$\widehat{\mathbf{x}}$
output
profile
vector

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 $\tilde{\mathbf{x}}$. During readout, the sequence $\tilde{\mathbf{x}}$ 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.

Theorem 2.1. An $(N, d+1)$-AECC corrects any asymmetric error of weight at most d.

Next, we define the *ℓ-gram distance* between two words $\mathbf{x}, \mathbf{y} \in [\![q]\!]^n$ as

$$d_{\text{gram}}(\mathbf{x}, \mathbf{y}; q, \ell) \triangleq d_{\text{asym}}(\mathbf{p}(\mathbf{x}; q, \ell), \mathbf{p}(\mathbf{y}; q, \ell)).$$

Note that d_{gram} is not a metric, as $d_{\text{gram}}(\mathbf{x}, \mathbf{y}; \ell) = 0$ does not imply that $\mathbf{x} = \mathbf{y}$. For example, $d_{\text{gram}}(0010, 1001; 2, 2) = 0$. Nevertheless, $([\![q]\!]^n, d_{\text{gram}})$ forms a pseudometric space. We convert this space into a metric space via an equivalence relation called metric identification. Specifically, we say that $\mathbf{x} \overset{d_{\text{gram}}}{\sim} \mathbf{y}$ if and only if $d_{\text{gram}}(\mathbf{x}, \mathbf{y}; q, \ell) = 0$. Then, by defining $\mathcal{Q}(n; q, \ell) \triangleq [\![q]\!]^n / \overset{d_{\text{gram}}}{\sim}$, we can make $(\mathcal{Q}(n; q, \ell), d_{\text{gram}})$ into a metric space.

Suppose that our data is encoded by a vector in $\mathbf{x} \in [\![q]\!]^n$ and let $\hat{\mathbf{x}}$ be the channel output profile. In what follows, we characterize the properties of the error vector $\mathbf{e} \triangleq \mathbf{p}(\mathbf{x}; q, \ell) - \hat{\mathbf{x}}$.

(i) **Substitution errors**. Here, certain symbols in the word $\mathbf{x}$ may be changed as a result of erroneous synthesis. If one symbol is changed, in the perfect coverage case, ℓ ℓ-grams will decrease their values by one and ℓ ℓ-grams will increase their values by one. Hence, the error resulting from s substitutions equals $\mathbf{e} = \mathbf{e}_- - \mathbf{e}_+$, where $\mathbf{e}_+, \mathbf{e}_- \geq \mathbf{0}$ both have weight $s\ell$.

(ii) **Undersampling errors**. Such errors occur when not all ℓ-grams are sequenced. For example, suppose that $\mathbf{x} = 00000$, and that $\hat{\mathbf{x}}$ is the channel output 3-gram vector. Undersampling of one 3-gram results in the weight of $\hat{\mathbf{x}}_{000}$ being four instead of five. Note that undersampling of t ℓ-grams results in an asymmetric error $\mathbf{e} \geq \mathbf{0}$ of weight t.

Let $\mathcal{C} \subseteq \mathcal{Q}(n; q, \ell)$. If $d = \min\{d_{\text{gram}}(\mathbf{x}, \mathbf{y}; \ell) : \mathbf{x}, \mathbf{y} \in \mathcal{C}, \mathbf{x} \neq \mathbf{y}\}$, then $\mathcal{C}$ is called an $(n, d; q, \ell)$ *ℓ-gram reconstruction code (GRC)*.

Proposition 2.2. An $(n, d; q, \ell)$-GRC can correct s substitution errors and t undersampling errors if $d > 2s\ell + t$.

Proof: Consider an $(n, d; q, \ell)$-GRC $\mathcal{C}$ and the set $\mathbf{p}(\mathcal{C}) = \{\mathbf{p}(\mathbf{x}) : \mathbf{x} \in \mathcal{C}\}$. By construction, $\mathbf{p}(\mathcal{C})$ is an (N, d)-AECC with $N = q^\ell$ that corrects all asymmetric errors of weight $\leq 2s\ell + t$. Suppose that on the contrary that $\mathcal{C}$ cannot correct s substitution errors and t errors due to undersampling. Then, there exists $\mathbf{x}, \mathbf{x}' \in C$ and error vectors $\mathbf{e}_{s,+}, \mathbf{e}_{s,-}, \mathbf{e}_t, \mathbf{e}'_{s,+}, \mathbf{e}'_{s,-}, \mathbf{e}'_t$ such that

$$\hat{\mathbf{x}} + \mathbf{e}_{s,+} - \mathbf{e}_{s,-} - \mathbf{e}_t = \hat{\mathbf{x}}' + \mathbf{e}'_{s,+} - \mathbf{e}'_{s,-} - \mathbf{e}'_t.$$

Here, $\mathbf{e}_{s,-} - \mathbf{e}_{s,+}$ and $\mathbf{e}'_{s,-} - \mathbf{e}'_{s,+}$ are the error vectors due to substitutions in $\mathbf{x}$ and $\mathbf{x}'$, respectively; each of the vectors $\mathbf{e}_{s,-}, \mathbf{e}_{s,+}, \mathbf{e}'_{s,-}, \mathbf{e}'_{s,+}$ has weight $s\ell$; $\mathbf{e}_t$ and $\mathbf{e}'_t$ are the undersampling error vectors of $\mathbf{x}$ and $\mathbf{x}'$, respectively, and both $\mathbf{e}_t, \mathbf{e}'_t$ have weight t. Therefore,

$$\hat{\mathbf{x}} - (\mathbf{e}_{s,-} + \mathbf{e}_t + \mathbf{e}'_{s,+}) = \hat{\mathbf{x}}' - (\mathbf{e}'_{s,-} + \mathbf{e}'_t + \mathbf{e}_{s,+}),$$

where $\mathbf{e}_{s,-} + \mathbf{e}_t + \mathbf{e}'_{s,+}$ and $\mathbf{e}'_{s,-} + \mathbf{e}'_t + \mathbf{e}_{s,+}$ are nonnegative vectors of weight at most $2s\ell + t$. This contradicts the fact that $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}'$ belong to a code that corrects asymmetric errors with weight at most $2s\ell + t$. ■

3. Upper Bounds on Code Size

In this section, we fix q and ℓ and first focus on determining the largest possible size of an $(n, d = 1; q, \ell)$-GRC. Equivalently, we determine the size of $\mathcal{Q}(n; q, \ell)$ – i.e., enumerate the distinct profile vectors of words in $[\![q]\!]^\ell$. We then proceed to describe an upper bound for general $(n, d; q, \ell)$-GRCs.

Instead of working with $\mathcal{Q}(n; q, \ell)$ directly, we consider only *closed* words, that is, words that start and end with the same $(\ell - 1)$-gram. Henceforth, $\bar{\mathcal{Q}}(n; q, \ell)$ denotes the set of closed words with distinct profile vectors.

We start with a discussion of the results derived by Jacquet *et al.* [9]. In this setting, consider the family of de Bruijn digraphs [11]. Given q and ℓ, the *de Bruijn digraph* $D(q, \ell)$ is defined on the vertex set $[\![q]\!]^{\ell-1}$. The pair $(\mathbf{w}, \mathbf{w}')$ belongs to the arc set if and only if $w_i = w'_{i-1}$ for $2 \leq i \leq \ell$. Hence, we identify the arc set with $[\![q]\!]^\ell$. In addition, for a vertex $\mathbf{w} \in [\![q]\!]^{\ell-1}$, we use $E^+(\mathbf{w})$ to denote the set of its q outgoing arcs and $E^-(\mathbf{w})$ to denote the set of its q incoming arcs.

Suppose that $\mathbf{u} = (u_{\mathbf{z}})_{\mathbf{z} \in [\![q]\!]^\ell}$ is a profile vector of some q-ary word of length n. Then the following $q^{\ell-1}$ equations, that we refer to as the *flow conservation equations*, hold true:

$$\sum_{\mathbf{z} \in E^+(\mathbf{w})} u_{\mathbf{z}} = \sum_{\mathbf{z} \in E^-(\mathbf{w})} u_{\mathbf{z}} \quad \text{for all } \mathbf{w} \in [\![q]\!]^{\ell-1}. \tag{1}$$

Since the number of ℓ-grams is $n - \ell + 1$, we also have

$$\sum_{\mathbf{z} \in [\![q]\!]^\ell} u_{\mathbf{z}} = n - \ell + 1. \tag{2}$$

Let $\mathbf{A}_{\text{gram}}$ be the incidence matrix of $D(q,\ell)$, augmented with a top row of all-ones; let $\mathbf{b}$ be a vector of length $q^{\ell-1}+1$ with a one as its first entry, and zeros elsewhere. Equations (1) and (2) may then be rewritten as $\mathbf{A}_{\text{gram}}\mathbf{u}=(n-\ell+1)\mathbf{b}$.

Define next the polytope

$$\mathcal{P}_{\text{gram}} \triangleq \{\mathbf{u}\in\mathbb{R}_{\geq 0}^{q^\ell} : \mathbf{A}_{\text{gram}}\mathbf{u}=\mathbf{b}\}, \tag{3}$$

and the corresponding sets of integer solutions

$$\mathcal{F}(n;q,\ell) \triangleq \{\mathbf{u}\in\mathbb{Z}_{\geq 0}^{q^\ell} : \mathbf{A}_{\text{gram}}\mathbf{u}=(n-\ell+1)\mathbf{b}\}, \tag{4}$$

$$\mathcal{E}(n;q,\ell) \triangleq \{\mathbf{u}\in\mathbb{Z}_{>0}^{q^\ell} : \mathbf{A}_{\text{gram}}\mathbf{u}=(n-\ell+1)\mathbf{b}\}. \tag{5}$$

The preceding discussion asserts that any profile vector must lie in $\mathcal{F}(n;q,\ell)$. Conversely, observe that any vector in $\mathcal{E}(n;q,\ell)$ has positive entries and so, we have the following lemma.

Lemma 3.1. If $\mathbf{u}\in\mathcal{E}(n;q,\ell)$, then there exists a q-ary word $\mathbf{x}$ of length n such that $\mathbf{p}(\mathbf{x};q,\ell)=\mathbf{u}$.

Proof: Construct a multidigraph D' on the vertex set $[\![q]\!]^{\ell-1}$ by adding $u_{\mathbf{z}}$ copies of the arc $\mathbf{z}$, for all $\mathbf{z}\in[\![q]\!]^{\ell}$. Since each $u_{\mathbf{z}}$ is positive, D' is strongly connected. Since $\mathbf{u}\in\mathcal{E}(n;q,\ell)$, $\mathbf{u}$ also satisfies the flow conservation equations and D' is consequently Eulerian. Also, D' has $n-\ell+1$ arcs and an Eulerian circuit on D' yields one such desired word $\mathbf{x}$. ■

Therefore, we have the following inequality,

$$|\mathcal{E}(n;q,\ell)| \leq |\bar{\mathcal{Q}}(n;q,\ell)| \leq |\mathcal{F}(n;q,\ell)| \ . \tag{6}$$

Jacquet *et al.* showed that these three quantities are asymptotically equivalent. Then, using techniques from analytic combinatorics, they also determined the following result on the asymptotic size for $\bar{\mathcal{Q}}(n;q,\ell)$. Note that $f(n)\sim g(n)$ stands for $\lim_{n\to\infty} f(n)/g(n)=1$.

Theorem 3.2 (Jacquet *et al.* [9]). Fix q,ℓ to be constant. Let $\mathcal{E}(n;q,\ell)$, $\mathcal{F}(n;q,\ell)$, $\mathcal{Q}(n,q,\ell)$ and $\bar{\mathcal{Q}}(n,q,\ell)$ be defined as above. Then

$$|\mathcal{E}(n;q,\ell)| \sim |\mathcal{F}(n;q,\ell)| \sim |\bar{\mathcal{Q}}(n,q,\ell)| \sim c(q,\ell)n^{q^\ell-q^{\ell-1}},$$

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

As hinted by (3) to (5), to enumerate codewords of interest, we need to enumerate lattice points in certain polytopes. This 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.* [12], which is based on an algorithm of Barvinok [13].

In the remainder of this section, we introduce fundamental ideas of Ehrhart theory and re-derive the results of [9] within this framework. We proceed to extend the enumeration procedure to profiles at a prescribed distance.

A. Ehrhart Theory

The first general treatment of the theory of enumerating lattice points in polytopes was described by Ehrhart [14], but we follow the combinatorial treatment of Beck and Robins [15].

Consider any *rational polytope* $\mathcal{P}$ given by

$$\mathcal{P} \triangleq \{\mathbf{u}\in\mathbb{R}^n_{\geq 0} : \mathbf{A}\mathbf{u}=\mathbf{b}\},$$

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

$$L_{\mathcal{P}}(t) \triangleq \#\{\mathbf{u}\in\mathbb{Z}^n_{\geq 0} : \mathbf{A}\mathbf{u}=t\mathbf{b}\}, \text{ for all } t\in\mathbb{Z}_{>0}.$$

Ehrhart [14] introduced the lattice point enumerator for rational polytopes and showed that $L_{\mathcal{P}}(t)$ is a “polynomial” in t whose “coefficients” are periodic in t. Formally, we define a *quasipolynomial* 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 in t. If c_D is not identically equal to zero, the quasipolynomial is said to be of *degree* D.

Furthermore, Ehrhart showed that the degree of $L_{\mathcal{P}}(t)$ corresponds to 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 $\mathcal{P}$. We provide next a formal statement of Ehrhart theorem.

Theorem 3.3 (Ehrhart theorem for polytopes [15, 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. Furthermore, if $\mathcal{P}$ is integer, then $L_{\mathcal{P}}(t)$ is a polynomial of degree D.

Motivated by (5), we consider the *interior* of $\mathcal{P}$ given by

$$\mathcal{P}^\circ \triangleq \{\mathbf{u}\in\mathcal{P} : \text{ for all } \mathbf{u}'\in\mathcal{P}, \text{ there exists an } \epsilon>0 \text{ such that } \mathbf{u}+\epsilon(\mathbf{u}-\mathbf{u}')\in\mathcal{P}\},$$

and for a positive integer t, we consider the quantity

$$L_{\mathcal{P}^\circ}(t)=\#\{\mathbf{u}\in\mathbb{Z}^n_{>0} : \mathbf{A}\mathbf{u}=t\mathbf{b}\}.$$

Ehrhart conjectured the following relation between $L_{\mathcal{P}}(t)$ and $L_{\mathcal{P}^\circ}(t)$, which was proved by Macdonald [16].

Theorem 3.4 (Ehrhart-Macdonald reciprocity [15, 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

$$L_{\mathcal{P}}(-t)=(-1)^D L_{\mathcal{P}^\circ}(t).$$

From the definition of lattice point enumerators, we may write $|\mathcal{F}(n;q,\ell)|=L_{\mathcal{P}_{\text{gram}}}(n-\ell+1)$. It can be shown (the proof is delegated to the full paper) that the polytope $\mathcal{P}_{\text{gram}}$ has dimension $q^\ell-q^{\ell-1}$ and $|\mathcal{E}(n;q,\ell)|=L_{\mathcal{P}^\circ_{\text{gram}}}(n-\ell+1)$. Hence, $L_{\mathcal{P}_{\text{gram}}}(n-\ell+1)$ and $L_{\mathcal{P}^\circ_{\text{gram}}}(n-\ell+1)$ are both quasipolynomials of degree $q^\ell-q^{\ell-1}$ in n. Furthermore, Theorem 3.2 implies that the leading coefficients of these quasipolynomials are in fact the same and constant (i.e., aperiodic). Using `LattE`, we can compute these leading coefficients for various pairs of values (q,ℓ). 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.* [9]), numerical computations of $c(q,\ell)$ are feasible for moderate values of (q,ℓ).

We conclude the section with an upper bound for the size of GRC codes for general d. The proof of the lemma is omitted due to space constraints.

Lemma 3.5. Let $C(n,d;q,\ell)$ denote the maximum size of a $(n,d;q,\ell)$-GRC, and let $t=\lfloor (d-1)/2 \rfloor$. Then

$$C(n,d;q,\ell) \leq \left(\frac{n-\ell+2}{t+1}\right)^{q^{\ell}} (q^{\ell})! \,.$$

Observe that for the case $d=1$ (i.e., $t=0$), and for fixed values of q and ℓ, Lemma 3.5 reduces to $C(n,d;q,\ell)=O(n^{q^{\ell}})$. On the other hand, the previous derivations imply that $C(n,1;q,\ell)=O(n^{q^{\ell}-q^{\ell-1}})$, improving the upper bound for this particular instance.

4. Constructive Lower Bounds

Fix q,ℓ,n and let $\mathbf{p}\mathcal{Q}(n;q,\ell)$ denote the set of all ℓ-gram profile vectors of words in $[\![q]\!]^{\ell}$. For ease of exposition, we identify words in $\mathcal{Q}(n;q,\ell)$ with their corresponding profile vectors in $\mathbf{p}\mathcal{Q}(n;q,\ell)$ and refer to GRCs as being subsets of $\mathcal{Q}(n;q,\ell)$ or $\mathbf{p}\mathcal{Q}(n;q,\ell)$ interchangeably.

To construct GRCs from known AECCs, we proceed as follows. Suppose that $\mathcal{C}$ is an $(N,d+1)$-AECC, where $N=q^{\ell}$. Then $\mathcal{C}\cap\mathbf{p}\mathcal{Q}(n;q,\ell)$ is an $(n,d+1;q,\ell)$-GRC and the objective of this section is to estimate $|\mathcal{C}\cap\mathbf{p}\mathcal{Q}(n;q,\ell)|$.

For this purpose, we use a classical family of AECCs proposed by Varshamov [17]. Fix d and let p be a prime such that $p>d$ and $p>N$. Choose n distinct nonzero elements $\alpha_1,\alpha_2,\ldots,\alpha_N$ in $\mathbb{Z}/p\mathbb{Z}$ and consider the matrix

$$\mathbf{H} \triangleq \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_N \\ \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^d & \alpha_2^d & \cdots & \alpha_N^d \end{pmatrix}.$$

Pick any vector $\boldsymbol{\beta}\in(\mathbb{Z}/p\mathbb{Z})^N$ and define the code

$$\mathcal{C}(\mathbf{H},\boldsymbol{\beta}) \triangleq \{\mathbf{u} : \mathbf{H}\mathbf{u} \equiv \boldsymbol{\beta} \bmod p\}.$$

It can be shown that $\mathcal{C}(\mathbf{H},\boldsymbol{\beta})$ is an $(N,d+1)$-AECC [17]. Hence, $\mathcal{C}(\mathbf{H},\boldsymbol{\beta})\cap\mathbf{p}\mathcal{Q}(n;q,\ell)$ is an $(n,d+1;q,\ell)$-GRC for all $\boldsymbol{\beta}\in(\mathbb{Z}/p\mathbb{Z})^N$. Therefore, by the pigeonhole principle, there exists a $\boldsymbol{\beta}$ such that $|\mathcal{C}(\mathbf{H},\boldsymbol{\beta})\cap\mathbf{p}\mathcal{Q}(n;q,\ell)|$ is at least $|\mathbf{p}\mathcal{Q}(n;q,\ell)|/p^d$. However, the choice of $\boldsymbol{\beta}$ that guarantees this lower bound is unclear.

In the rest of this section, we fix a certain choice of $\mathbf{H}$ and $\boldsymbol{\beta}$ and provide lower bounds on the size of $\mathcal{C}(\mathbf{H},\boldsymbol{\beta})\cap\mathbf{p}\mathcal{Q}(n;q,\ell)$ as a function of n. As before, instead of looking at $\mathbf{p}\mathcal{Q}(n;q,\ell)$ directly, we consider the set of closed words $\bar{\mathcal{Q}}(n;q,\ell)$ and the corresponding set of profile vectors $\mathbf{p}\bar{\mathcal{Q}}(n;q,\ell)$.

Let $\boldsymbol{\beta}=\mathbf{0}$ and choose $\mathbf{H}$ and p so that $\mathbf{1}$ belongs to the $(N,d+1)$-AECC $\mathcal{C}(\mathbf{H},\mathbf{0})$. In other words, if we regard $\mathbf{H}$ as a matrix over positive integers, we have $\mathbf{H}\mathbf{1}=p\boldsymbol{\beta}'$ for some $\boldsymbol{\beta}'\geq 0$.

Define the $(q^{\ell-1}+1+d)\times(q^{\ell}+d)$-matrix

$$\mathbf{A}(\mathbf{H},p) \triangleq \left(\begin{array}{c|c} \mathbf{A}_{\mathrm{gram}} & \mathbf{0} \\ \hline \mathbf{H} & -p\mathbf{I}_d \end{array}\right),$$

where $\mathbf{A}_{\mathrm{gram}}$ is defined as in Section 3. Let $\mathbf{b}$ be a vector of length $q^{\ell-1}+1+d$ that has one as the first entry and zeros elsewhere.

The following proposition demonstrates that $|\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathcal{E}(n;q,\ell)|$ is given by the number of lattice points in the interior of a certain polytope.

Proposition 4.1. Let $\mathcal{C}(\mathbf{H},\mathbf{0})$, $\mathbf{A}(\mathbf{H},p)$ and $\mathbf{b}$ be defined as above. Then $\#\{\mathbf{u}:\mathbf{A}(\mathbf{H},p)\mathbf{u}=(n-\ell+1)\mathbf{b}$ and $\mathbf{u}>\mathbf{0}\}=|\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathcal{E}(n;q,\ell)|$.

Proof: Let $\mathbf{u}>\mathbf{0}$ be such that $\mathbf{A}(\mathbf{H},p)\mathbf{u}=(n-\ell+1)\mathbf{b}$. Consider the vector $\mathbf{u}_0$ which equals the vector $\mathbf{u}$ restricted to the first N coordinates. Then $\mathbf{A}_{\mathrm{gram}}\mathbf{u}_0=(n-\ell+1)\mathbf{b}_0$, where $\mathbf{b}_0$ is a vector of length $q^{\ell-1}+1$ with one at the first coordinate and zeros elsewhere. So, $\mathbf{u}_0\in\mathcal{E}(n;q,\ell)$. On the other hand, $\mathbf{H}\mathbf{u}_0=p\boldsymbol{\beta}'$ for some $\boldsymbol{\beta}'>0$. In other words, $\mathbf{H}\mathbf{u}_0\equiv\mathbf{0} \bmod p$ and so $\mathbf{u}_0\in\mathcal{C}(\mathbf{H},\mathbf{0})$.

Therefore, $\mathbf{u}\mapsto\mathbf{u}_0$ is a map from $\{\mathbf{u}:\mathbf{A}(\mathbf{H},p)\mathbf{u}=(n-\ell+1)\mathbf{b}$ and $\mathbf{u}>\mathbf{0}\}$ to $\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathcal{E}(n;q,\ell)$. It can be verified that this map is a bijection. ■

Next, we provide two technical lemmas whose proofs are omitted due to space limitations. The first lemma justifies our choice of $\mathbf{H}$ and p.

Lemma 4.2. Let $\mathcal{C}(\mathbf{H},\mathbf{0})$, $\mathbf{A}(\mathbf{H},p)$ and $\mathbf{b}$ be defined as above. If $\mathbf{1}\in\mathcal{C}(\mathbf{H},\mathbf{0})$, then the polytope given by $\{\mathbf{u}\in\mathbb{R}^N:\mathbf{u}\geq\mathbf{0}$ and $\mathbf{A}(\mathbf{H},p)\mathbf{u}=(n-\ell+1)\mathbf{b}\}$ has dimension $q^{\ell}-q^{\ell-1}$.

To simplify our arguments, we consider values of $n-\ell+1$ that ensure the polytope is integer. Specifically, we choose n so that $n-\ell+1$ belongs to a certain congruence class.

Lemma 4.3. Let $\lambda=\mathrm{lcm}(1,2,\ldots,q^{\ell-1},p)$ and let $\mathbf{b}'=\lambda\mathbf{b}$. Then

$$\mathcal{P}_{\mathrm{GRC}} \triangleq \{\mathbf{u}\in\mathbb{R}^N_{\geq 0} : \mathbf{A}(\mathbf{H},p)\mathbf{u}=\mathbf{b}'\} \tag{7}$$

is an integer polytope.

Therefore, Lemmas 4.2, 4.3 and Ehrhart's theorem for integer polytopes imply that the lattice point enumerator $L_{\mathcal{P}_{\mathrm{GRC}}}(t)$ is a polynomial of degree $q^{\ell}-q^{\ell-1}$. Furthermore, the leading coefficient of $L_{\mathcal{P}_{\mathrm{GRC}}}(t)$ provides a lower bound on $|\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathbf{p}\mathcal{Q}(n;q,\ell)|$.

Theorem 4.4. Fix q,ℓ,d. Choose $\mathbf{H}$ and p such that $\mathcal{C}(\mathbf{H},\mathbf{0})$ is an $(N,d+1)$-AECC containing $\mathbf{1}$. Let $\mathcal{P}_{\mathrm{GRC}}$ be the polytope defined by (7) with $\lambda=\mathrm{lcm}(1,2,\ldots,q^{\ell-1},p)$. Suppose its lattice point enumerator $L_{\mathcal{P}_{\mathrm{GRC}}}(t)$ has leading coefficient $c(q,\ell,d)$. Then

$$\limsup_{n\to\infty}\frac{|\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathbf{p}\mathcal{Q}(n;q,\ell)|}{c(q,\ell,d)(n/\lambda)^{q^{\ell}-q^{\ell-1}}}\geq 1.$$

Proof: For any positive integer t, let $n-\ell+1=t\lambda$. Then $|\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathbf{p}\mathcal{Q}(n;q,\ell)|\geq|\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathbf{p}\bar{\mathcal{Q}}(n;q,\ell)|\geq L_{\mathcal{P}^{\circ}_{\mathrm{GRC}}}(t)$. Since $L_{\mathcal{P}_{\mathrm{GRC}}}(t)$ is a polynomial of degree $q^{\ell}-q^{\ell-1}$, by Ehrhart-Macdonald reciprocity, $L_{\mathcal{P}^{\circ}_{\mathrm{GRC}}}(t)$ is also a polynomial of the same degree with the same leading coefficient. Hence, $|\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathbf{p}\mathcal{Q}(n;q,\ell)|\geq c(q,\ell,d)t^{q^{\ell}-q^{\ell-1}}+O(t^{q^{\ell}-q^{\ell-1}-1})=c(q,\ell,d)((n-\ell+1)/\lambda)^{q^{\ell}-q^{\ell-1}}+O(t^{q^{\ell}-q^{\ell-1}-1})$. Taking limits yields the desired result. ■

TABLE I

COMPUTATIONS OF LATTICE POINT ENUMERATORS FOR $\mathcal{P}^{\circ}_{\rm GRC}$

When $(q,\ell)=(2,3)$, we have $c(2,3)=1/288$. Here, $t=(n-\ell+1)/\lambda$.

d	p	λ	$c(2,3,d)$	$c(2,3,d)/\lambda^4$	$c(2,3)/p^d$	$L_{\mathcal{P}^{\circ}_{\rm GRC}}(t)$
1	11	132	95832	1/3168	1/3168	$95832\,t^4-11616\,t^3+517\,t^2-10\,t+1$
2	13	156	12168	1/48672	1/48672	$12168\,t^4-1248\,t^3+131\,t^2-16\,t+1$
3	13	156	936	1/632736	1/632736	$936\,t^4-96\,t^3+107\,t^2-10\,t+1$
4	17	204	72	1/24054048	1/24054048	$72\,t^4-96\,t^3+47\,t^2-10\,t+1$
5	17	204	72	1/24054048	1/408918816	$72\,t^4-96\,t^3+47\,t^2-10\,t+1$
6	17	204	72	1/24054048	1/6951619872	$72\,t^4-96\,t^3+47\,t^2-10\,t+1$

Theorem 4.4 guarantees that the code size is at least $c(q,\ell,d)(n/\lambda)^{q^\ell-q^{\ell-1}}$, where $c(q,\ell,d)$ is the leading coefficient of $L_{\mathcal{P}_{\rm GRC}}(t)$.

Remark 1. We provide an alternative view of the problem of enumerating the codewords in $\mathcal{C}(H,\mathbf{0})\cap\mathcal{E}(n;q,\ell)$. Let $\mathbf{A}_0$ be the $(q^{\ell-1}+d)\times(q^\ell+d)$-matrix that results from removing the first row of $\mathbf{A}(\mathbf{H},p)$. Consider the code $\mathcal{C}_0\triangleq\{\mathbf{u}\in\mathbb{Z}^N_{\geq 0}:\mathbf{A}_0(\mathbf{u},\boldsymbol{\beta}')=\mathbf{0}$ for some $\boldsymbol{\beta}'\in\mathbb{Z}^d\}$. Then $\mathcal{C}_0\subset\mathcal{C}(\mathbf{H},\mathbf{0})$ and is therefore an $(N,d+1)$-AECC. Furthermore, any vector in $\mathcal{C}_0$ satisfies the flow conservation equations (1).

If we consider all codewords of L_1-weight $(n-\ell-1)$ in $\mathcal{C}_0$, we obtain profile vectors of words with length n. In other words, our enumeration problem is related to the problem of determining the *weight enumerator* of the code $\mathcal{C}_0$. We remark that the weight enumerator of the classical Varshamov codes has been studied by Stanley and Yoder [18] and Delsarte and Piret [19]. Properties of this weight enumerator have also been implicitly used by Graham and Sloane to construct binary constant weight codes [20].

A. Computational Results

By Proposition 4.1, we have $|\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathcal{E}(n;q,\ell)|=L_{\mathcal{P}^{\circ}_{\rm GRC}}(t)$ whenever $n-\ell+1=t\lambda$. By Ehrhart-Macdonald reciprocity, this polynomial is given by $L_{\mathcal{P}^{\circ}_{\rm GRC}}(t)=(-1)^{q^\ell-q^{\ell-1}}L_{\mathcal{P}_{\rm GRC}}(-t)$. Using LattE, we determined $L_{\mathcal{P}_{\rm GRC}}(t)$, $L_{\mathcal{P}^{\circ}_{\rm GRC}}(t)$ as well as $c(q,\ell,d)$ for certain values of q, ℓ and d. The results are summarized in Table I.

Example 4.1. Let $(q,\ell,d)=(2,3,2)$. Choose $p=13$ and

$$\mathbf{H}=\begin{pmatrix}1&2&3&5&8&10&11&12\\1&4&9&12&12&9&4&1\end{pmatrix}.$$

Then $\mathcal{C}(\mathbf{H},\mathbf{0})$ is an $(8,3)$-AECC and contains the all-one vector. Here, $\lambda=156$ and for $n=156t+2$, the number of codewords in $\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathcal{E}(n;2,3)$ is given by

$$L_{\mathcal{P}^{\circ}_{\rm GRC}}(t)=12168\,t^4-1248\,t^3+131\,t^2-16\,t+1.$$

When $t=1$ or $n=158$, there exist a $(158,3;2,3)$-GRC of size 11036.

We conclude with a conjecture on the relation between $c(q,\ell)$ and $c(q,\ell,d)$.

Conjecture 4.5. Fix q,ℓ,d. Choose $\mathbf{H}$ and p such that $\mathcal{C}(\mathbf{H},\mathbf{0})$ is an $(N,d+1)$-AECC containing $\mathbf{1}$. Let $c(q,\ell)$, $c(q,\ell,d)$ and λ be the quantities defined in Theorems 3.2 and 4.4. Then $c(q,\ell,d)/\lambda^{q^\ell-q^{\ell-1}}\geq c(q,\ell)/p^d$.

Roughly speaking, the conjecture states that asymptotically, $|\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathcal{E}(n;q,\ell)|$ is at least $|\bar{\mathcal{Q}}(n;q,\ell)|/p^d$. In other words, for this particular choice of $\mathbf{H}$ and $\boldsymbol{\beta}$, we asymptotically achieve the code size guaranteed by the pigeonhole principle.

REFERENCES

[1] S. Kannan and A. McGregor, "More on reconstructing strings from random traces: insertions and deletions," in *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on.* 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 *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on.* IEEE, 2010, pp. 1238–1242.

[3] ——, "Quadratic-backtracking algorithm for string reconstruction from substring compositions," in *Information Theory (ISIT), 2014 IEEE International Symposium on.* 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] 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.

[10] T. Kløve, *Error correcting codes for the asymmetric channel*. Department of Pure Mathematics, University of Bergen, 1981.

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

[12] 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.

[13] 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.

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

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

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

[17] R. Varshamov, "A class of codes for asymmetric channels and a problem from the additive theory of numbers," *IEEE Trans. Inform. Theory*, vol. 19, no. 1, pp. 92–95, 1973.

[18] R. P. Stanley and M. F. Yoder, "A study of varshamov codes for asymmetric channels," *Jet Prop. Lab. Tech. Rep*, pp. 32–1526, 1972.

[19] P. Delsarte and P. Piret, "Spectral enumerators for certain additive-error-correcting codes over integer alphabets," *Information and Control*, vol. 48, no. 3, pp. 193–210, 1981.

[20] R. L. Graham and N. J. A. Sloane, "Lower bounds for constant weight codes," *IEEE Trans. Inform. Theory*, vol. 24, pp. 70–75, 1980.