IEEE TRANSACTIONS ON
INFORMATION THEORY

Codes for DNA Sequence Profiles

Han Mao Kiah*, Gregory J. Puleo†, and Olgica Milenkovic†

***Abstract*— We consider the problem of storing and retrieving information from synthetic DNA media. We introduce the DNA storage channel and model the read process through the use of profile vectors. We provide an asymptotic analysis of the number of profile vectors and propose new asymmetric coding techniques to combat the effects of synthesis and sequencing noise. Furthermore, we construct two families of codes for this new channel model.**

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 [3]–[5]. With the recent development of archival DNA-based storage devices [6], [7] and rewritable, random-access DNA storage media [8], 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 write and read errors. The write process in DNA-based storage systems is DNA synthesis, a biochemical process that allows for creating moderately long DNA strings via the use of columns or microarrays [9]. Synthesis involves sequential inclusion of bases into a growing string, and is accompanied by chemical error correction. The read process in DNA-based storage is DNA sequencing, while classical decoding is replaced by a combination of assembly and error-control decoding. DNA sequencing operates by creating many copies of the same string and then fragmenting them into a collection of substrings (reads) of approximately the same length, ℓ, so as to produce a large number of overlapping "reads". The larger the number of sequence replicas and reads, the larger the *coverage* of the sequence – the average number of times a symbol in the sequence is contained in a read. Assembly aims to reconstruct the original sequence by stitching the overlapping fragments together; the assembly procedure is NP-hard under most formulations [10]. Nevertheless, practical approximation algorithms based on Eulerian paths in de Bruijn graphs have shown to offer good reconstruction performance under high-coverage [11]. Due to the high cost of synthesis, most current DNA storage systems do not use sequence lengths n exceeding several thousands nucleotides (nts). Synthesis error rates range between 0.1 and 3% depending on the cost of the technology [9], [12], and the errors are predominantly substitution errors. The read length ℓ typically ranges anywhere between 100 to 1500 nts, although some technologies even produce reads of lengths exceeding 10, 000 nts. 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% [13]. In the former case, the dominant error events are substitution errors [14]. Furthermore, due to non-uniform fragmentation, some proper substrings are not available during the reading stage, leaving what is known as coverage gaps in the original message.

More formally, to store and retrieve information in DNA one starts with a desired information sequence encoded into a sequence $\mathbf{x} \in \mathcal{D} = \{A, T, G, C\}^n$, where $\mathcal{D}$ denotes the nucleotide alphabet. The *DNA storage channel*, shown in Fig. 1 and formally defined in Section 2, models a physical process which takes as its input the sequence $\mathbf{x}$ of length n, and synthesizes (writes) it physically into a macromolecule string, denoted by $\widetilde{\mathbf{x}}$. Hence, DNA both encodes information and serves as a storage media. Ideally, one would like to synthesize $\mathbf{x}$ without errors, which is not possible in practice. As a result, the sequence $\widetilde{\mathbf{x}}$ is a distorted version of $\mathbf{x}$ in so far as it contains s_{syn} substitution errors, where s_{syn} is an integer value governed by the synthesis technology. When a user desires to retrieve the information, the process proceeds to amplify the string $\widetilde{\mathbf{x}}$ and then fragments all copies of the string, resulting in a highly redundant mix of reads. This mix may contain multiple copies of the same substring, say $\widetilde{\mathbf{x}}_1 = \widetilde{x}_1 \cdots \widetilde{x}_\ell$ as well as multiple copies of another substring $\widetilde{\mathbf{x}}_k = \widetilde{x}_k \cdots \widetilde{x}_{k+\ell-1}$, with $k \neq 1$ identical to $\widetilde{\mathbf{x}}_1$ (i.e., such that $\widetilde{\mathbf{x}}_1 = \widetilde{\mathbf{x}}_k$). 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 $\widetilde{\mathbf{x}}$. Hence, in the DNA storage channel we model the output of the fragmentation block as an *unordered subset of substrings (reads)* of the sequence $\widetilde{\mathbf{x}}$ of length ℓ, with $\ell < n$, denoted by $\widetilde{\mathcal{L}}(\mathbf{x}) = \{\widetilde{\mathbf{x}}_{i_1}, \ldots, \widetilde{\mathbf{x}}_{i_f}\}$, where $i_1 < i_2 < \ldots < i_f$, and where $f \leq n - \ell + 1$ is the number of reads. As an example, both $\widetilde{\mathbf{x}}_1$ and $\widetilde{\mathbf{x}}_k$ may be observed and hence included in the unordered set of substrings, or only one or neither. In the latter two cases, we say that the substring(s) were not covered during fragmentation.

Some of the observed substrings will contain additional substitution errors, due to the next step of sequencing or reading of the substrings. For simplicity, we assume that the total number of sequencing errors equals s_{seq}. The set of substrings at the output of the DNA storage channel is denoted by the multiset $\widehat{\mathcal{L}}(\mathbf{x}) = \{\widehat{\mathbf{x}}_{i_1}, \ldots, \widehat{\mathbf{x}}_{i_f}\}$, and each $\widehat{\mathbf{x}}_i$ may be a substitution-distorted version of $\widetilde{\mathbf{x}}_i$. The information contained in $\widehat{\mathcal{L}}(\mathbf{x})$ may be summarized by its multiplicity vector, also called *output profile vector* $\widehat{\mathbf{p}}(\mathbf{x})$, which is also our

This work was supported by the NSF STC Class 2010 CCF 0939370 and NSF CCF 1526875 grants, and the Strategic Research Initiative (SRI) Grant conferred by the University of Illinois, Urbana-Champaign. Research of the second author was supported in part by the IC Postdoctoral Research Fellowship. A limited sample of the results in this manuscript appeared in the Proceedings of the ITW 2015 [1] and the Proceedings of the ISIT 2015 [2].

Digital Object Identifier: 10.1109/TIT.2016.2555321

channel output. The profile vector is of length 4^ℓ, and each entry in the vector corresponds to exactly one of the ℓ-length strings over $\mathcal{D}$. The ordering of the ℓ-strings is assumed to be lexicographical. Furthermore, the jth entry in $\widehat{\mathbf{p}}(\mathbf{x})$ equals the number of times the j-th string in the lexicographical order was observed in $\widehat{\mathcal{L}}(\mathbf{x}) = \{\widehat{\mathbf{x}}_{i_1}, \ldots, \widehat{\mathbf{x}}_{i_f}\}$. Hence, for each $1 \leq j \leq 4^\ell$, the jth entry in $\widehat{\mathbf{p}}(\mathbf{x})$ is a value between 0 and $n - \ell + 1$.

The main contributions of the paper are as follows. The first contribution is to introduce the DNA storage channel and *model the read process (sequencing)* through the use of *profile vectors*. A profile vector of a sequence enumerates all substrings of the sequence, and profile vectors form a pseudometric space amenable for coding theoretic analysis[1]. The second contribution of the paper is to *introduce a new family of codes* for three classes of errors arising in the DNA storage channel due to synthesis, lack of coverage and sequencing, 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 (a) codewords with different profile vectors or profile vectors at sufficiently large distance from each other; and (b) 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 we allow arbitrary ℓ-substrings, the problem of enumerating all valid profile vectors was previously addressed by Jacquet *et al.* [15] in the context of "Markov types". However, the method of Jacquet *et al.* addressed Markov types which lead to substrings of length $\ell = 2$ only. Furthermore, the Markov type approach does not extend to the case of enumeration of profiles with specific ℓ-substring constraints or profiles at sufficiently large distance from each other, and hence the proof techniques used by the authors of [15] and those pursued in this work are substantially different.

We cast our more general enumeration and 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. We also describe two decoding procedures for sequence profiles that combine graph theoretical principles and sequencing by hybridization methods.

As our analysis involves tools from coding, graph theory and bioinformatics alike, many definitions and terms used may not be readily available in the standard information theory literature. To aid the reader, we have included a table of relevant definitions in Appendix A.

2. Profile Vectors and the DNA Storage Channel

We start this section by defining the relevant terminology and the DNA storage channel.

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* or a *substring* of $\mathbf{x}$ of length ℓ is a subsequence of $\mathbf{x}$ with ℓ consecutive indices. Let $\mathbf{p}(\mathbf{x}; q, \ell)$ denote the (ℓ-gram) *profile vector* of length q^ℓ, indexed by all words of $[\![q]\!]^\ell$ ordered lexicographically. We refer to the j-th word in this lexicographic order by $\mathbf{z}(j)$. 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 $\mathbf{x} \in [\![q]\!]^n$, the sum of entries in $\mathbf{p}(\mathbf{x}; q, \ell)$, equals $(n - \ell + 1)$.

For $\mathbf{x}, \mathbf{y} \in [\![q]\!]^n$, define the usual *Hamming distance* between a pair of words to be the number of coordinates where the two words differ. For $\mathbf{u}, \mathbf{v} \in \mathbb{Z}^N$, we define the *$L_1$-distance* between $\mathbf{u}$ and $\mathbf{v}$ to be the sum $\sum_{i=1}^N |u_i - v_i|$ and the *L_1-weight* of $\mathbf{u}$ to be the L_1-distance between $\mathbf{u}$ and $\mathbf{0}$. For brevity, the weight of a word stands for its L_1-weight.

Before we proceed with a formal definition of the DNA storage channel, we introduce the system errors that characterize such a channel. To this end, suppose that the data of interest is encoded by a vector $\mathbf{x} \in [\![q]\!]^n$ and let $\widehat{\mathbf{p}}(\mathbf{x})$ be the output profile of the DNA channel, as indicated in Fig. 1. The profile error vector, $\mathbf{e} \triangleq \mathbf{p}(\mathbf{x}; q, \ell) - \widehat{\mathbf{p}}(\mathbf{x})$ arises due to the following error events.

(i) **Substitution errors due to synthesis**. 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 counts by one and ℓ ℓ-grams will increase their counts by one. Hence, the error vector resulting from $s_{\rm syn}$ substitutions equals $\mathbf{e} = \mathbf{e}_- - \mathbf{e}_+$, where $\mathbf{e}_+, \mathbf{e}_-$ are vectors of weight $s_{\rm syn}\,\ell$ with $\mathbf{e}_+, \mathbf{e}_- \geq \mathbf{0}$.

(ii) **Coverage errors**. Such errors occur when not all ℓ-grams are observed during fragmentation and subsequently sequenced. For example, suppose that $\mathbf{x} = 00000$, and that $\widehat{\mathbf{p}}(\mathbf{x})$ is the channel output 3-gram profile vector. The coverage loss of one 3-gram results in the count of 000 in $\widehat{\mathbf{p}}(\mathbf{x})$ to be two instead of three. Note that imperfect coverage of t ℓ-grams results in an asymmetric error $\mathbf{e} \geq \mathbf{0}$ of weight t.

(iii) **ℓ-gram substitution errors due to sequencing**. Here, certain symbols in each fragment $\widetilde{\mathbf{x}}_i$ may be changed during the sequencing process. Suppose the ℓ-gram $\widetilde{\mathbf{x}}_i$ is altered to $\widehat{\mathbf{x}}_i$, $\widehat{\mathbf{x}}_i \neq \widetilde{\mathbf{x}}_i$. Then the count for $\widetilde{\mathbf{x}}_i$ will decrease by one while the count for $\widehat{\mathbf{x}}_i$ will increase by one. Hence, the error resulting from $s_{\rm seq}$ ℓ-gram substitutions equals $\mathbf{e} = \mathbf{e}_- - \mathbf{e}_+$, where $\mathbf{e}_+, \mathbf{e}_- \geq \mathbf{0}$, and $\mathbf{e}_+$ and $\mathbf{e}_-$ each has weight $s_{\rm seq}$.

Definition 2.1. The DNA storage channel with parameters $(n, q, \ell; t, s_{\rm syn}, s_{\rm seq})$ is a channel which takes as its input a vector $\mathbf{x} \in [\![q]\!]^n$ and outputs a vector $\widehat{\mathbf{p}}(\mathbf{x}) \in \mathbb{Z}^{q^\ell}$ such that there exists a $\widetilde{\mathbf{x}} \in [\![q]\!]^n$ and a vector $\widetilde{\mathbf{p}}(\mathbf{x}) \in \mathbb{Z}^{q^\ell}$ with the following properties:

(i) the Hamming distance between $\widetilde{\mathbf{x}}$ and $\mathbf{x}$ is at most $s_{\rm syn}$;
(ii) all entries of $\mathbf{p}(\widetilde{\mathbf{x}}; q, \ell) - \widetilde{\mathbf{p}}(\mathbf{x})$ are nonnegative and the L_1-weight of $\mathbf{p}(\widetilde{\mathbf{x}}; q, \ell) - \widetilde{\mathbf{p}}(\mathbf{x})$ is at most t;
(iii) there exist vectors $\mathbf{e}_-, \mathbf{e}_+ \geq 0$, each of weight at most

[1] A pseudometric space is a generalization of a metric space in which one allows the distance between two distinct points to be zero.

s_{seq}, such that $\widetilde{\mathbf{p}}(\mathbf{x}) = \widehat{\mathbf{p}}(\mathbf{x}) + \mathbf{e}_+ - \mathbf{e}_-$.

Here, properties (i)–(iii) correspond to the error types (i)–(iii) discussed before the definition.

Example 2.1. For simplicity, let $q = 2$, $\ell = 2$, $t = 1$, $s_{\text{syn}} = 1$, $s_{\text{seq}} = 2$, and assume that one wants to store the sequence $\mathbf{x} = 0110100$. One synthesis error, the maximum allowed under the given parameter constraints, will render $\mathbf{x}$ into a sequence $\widetilde{\mathbf{x}}$, say $\widetilde{\mathbf{x}} = 1110100$. The multiset of ℓ-grams belonging to $\widetilde{\mathbf{x}}$ is given by $\{11, 11, 10, 01, 10, 00\}$, and some of these ℓ-grams may be subjected to sequencing errors and possibly not observed due to coverage errors. Suppose that one copy of 10 is lost due to coverage errors, so that $\widetilde{\mathcal{L}}(\mathbf{x}) = \{11, 11, 10, 01, 00\}$, and that the second and third ℓ-grams are sequenced incorrectly, resulting in $\{11, 01, 11, 01, 00\}$. Hence, the DNA storage channel output would be the unordered set $\widehat{\mathcal{L}}(\mathbf{x}) = \{11, 01, 11, 01, 00\}$ which we summarize with the profile vector $\widehat{\mathbf{p}}(\mathbf{x}) = (1, 2, 0, 2)$. Note that none of the entries of $\widehat{\mathbf{p}}(\mathbf{x})$ exceeds $n - \ell + 1 = 6$, and that the sum of the entries equals five rather than six due to one coverage error.

Consider further a subset $S \subseteq [\![q]\!]^\ell$. For $\mathbf{x} \in [\![q]\!]^n$, we similarly define $\mathbf{p}(\mathbf{x}; S)$ to be the vector indexed by S, whose entry indexed by $\mathbf{z} \in S$ gives the number of occurrences of $\mathbf{z}$ as an ℓ-gram of $\mathbf{x}$. We are interested in vectors $\mathbf{x}$ whose ℓ-grams belong to S. Once again, the sum of entries in $\mathbf{p}(\mathbf{x}; S)$ equals $n - \ell + 1$.

The choice of S is governed by certain considerations in DNA sequence design, including:

(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 and overall coverage of the ℓ-grams. From the perspective of sequencing, GC contents of roughly 50% are desired[2].

To make this modeling assumption more precise and general, we assume sets S of the form described below. Suppose that $0 \le w_1 < w_2 \le \ell$ and $1 \le q^* \le q - 1$. Let $[w_1, w_2]$ denote the set of integers $\{w_1, w_1 + 1, \ldots, w_2\}$. For each $\mathbf{x} \in [\![q]\!]^\ell$, let the *$q^*$-weight* of $\mathbf{x}$ be the number of symbols in $\mathbf{x}$ that belong to $[q - q^*, q - 1]$, and denote the weight by $\mathrm{wt}(\mathbf{x}; q^*)$. Let

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

be the set of all sequences with q^* weights restricted to $[w_1, w_2]$. For example,

$$S(2, 4; 1, [2, 3])) = \{0011, 0101, 0110, 0111, 1001,\\ 1010, 1011, 1100, 1101, 1110\}.$$

We remark that if we represent A, T, G, C by $0, 1, 2, 3$, respectively, and set $q = 4$ and $q^* = 2$, the choice $w_1 = w_2 = \ell/2$ for even ℓ and the choices $w_1 = \lfloor \ell/2 \rfloor$ and $w_2 = w_1 + 1$ for odd ℓ enforce the balanced GC constraint. Also, note that $S(q, \ell; q^*, [0, \ell]) = [\![q]\!]^\ell$, for any choice of q^*.

(ii) **Forbidden ℓ-grams**. Studies have indicated that certain substrings in DNA sequences – such as GCG, CGC – are likely to cause sequencing errors (see [17]). 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.

Therefore, with an appropriate choice of S, we may lower the probability of substitution errors due to synthesis, lack of coverage and sequencing. Furthermore, as we show in our subsequent derivations, a carefully chosen set S may improve the error-correcting capability of a DNA-based storage system. This is achieved by designing codewords at a sufficiently large "distance" from each other and ensuring that the codewords avoid error-causing GC biases and substrings. Next, we formally define the notion of sequence and profile distance as well as error-correcting codes for the corresponding DNA channel.

3. Error-Correcting Codes for the DNA Storage Channel

Fix $S \subseteq [\![q]\!]^\ell$. Let N be an integer which usually denotes the number of ℓ-grams in the profile vector, i.e. $N = |S|$. Let $\mathbb{Z}_{\ge 0}^N$ denote the set of vectors of length N whose entries are nonnegative integers. For $\mathbf{u} \in \mathbb{Z}_{\ge 0}^N$, we sometimes write $\mathbf{u} \ge \mathbf{0}$. For any pair of words $\mathbf{u}, \mathbf{v} \in \mathbb{Z}_{\ge 0}^N$, let $\Delta(\mathbf{u}, \mathbf{v}) \triangleq \sum_{i=1}^N \max(u_i - v_i, 0)$ and define the *asymmetric distance* as $d_{\text{asym}}(\mathbf{u}, \mathbf{v}) = \max(\Delta(\mathbf{u}, \mathbf{v}), \Delta(\mathbf{v}, \mathbf{u}))$ [18]. A set $\mathcal{C}$ is called an *(N, d)-asymmetric error correcting code (AECC)* if $\mathcal{C} \subseteq \mathbb{Z}_{\ge 0}^N$ and $d = \min\{d_{\text{asym}}(\mathbf{x}, \mathbf{y}) : \mathbf{x}, \mathbf{y} \in \mathcal{C}, \mathbf{x} \ne \mathbf{y}\}$. For any $\mathbf{x} \in \mathcal{C}$, let $\mathbf{e} \in \mathbb{Z}_{\ge 0}^N$ be such that $\mathbf{x} - \mathbf{e} \ge \mathbf{0}$. 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 [18, Thm 9.1]).

Theorem 3.1. An $(N, d+1)$-AECC corrects any asymmetric error of L_1-weight at most d.

Next, we let $([\![q]\!]^n; S)$ denote all q-ary words of length n whose ℓ-grams belong to S and define the *ℓ-gram distance* between two words $\mathbf{x}, \mathbf{y} \in ([\![q]\!]^n; S)$ as

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

Note that d_{gram} is not a metric, as $d_{\text{gram}}(\mathbf{x}, \mathbf{y}; S) = 0$ does not imply that $\mathbf{x} = \mathbf{y}$. For example, we have $d_{\text{gram}}(0010, 1001; [\![2]\!]^2) = 0$. Nevertheless, $(([\![q]\!]^n; S), d_{\text{gram}})$ 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 $d_{\text{gram}}(\mathbf{x}, \mathbf{y}; S) = 0$. Then, by defining $\mathcal{Q}(n; S) \triangleq ([\![q]\!]^n; S)/\overset{d_{\text{gram}}}{\sim}$, we can make $(\mathcal{Q}(n; S), d_{\text{gram}})$ into a metric space. An element X in $\mathcal{Q}(n; S)$ is an equivalence class, where $\mathbf{x}, \mathbf{x}' \in X$ implies that $\mathbf{p}(\mathbf{x}; S) = \mathbf{p}(\mathbf{x}'; S)$. We specify the

[2] The reason behind the GC constraint is based on the observation that in Watson-Crick pairings, G and C bond with three, while A and T bond with two hydrogen bonds. Hence, the bonds between G and C are stronger, and having many stacked GC pairs or large GC content would make the DNA sequence more stable, but at the same time harder to fragment. It is known that GC rich substrings of DNA suffer most of the coverage errors during sequencing. On the other hand, a large AT content makes the DNA strand less stable and may cause occasional protrusions in DNA double helices. Hence, it is desirable to have a balance of GC bases in the string [16].

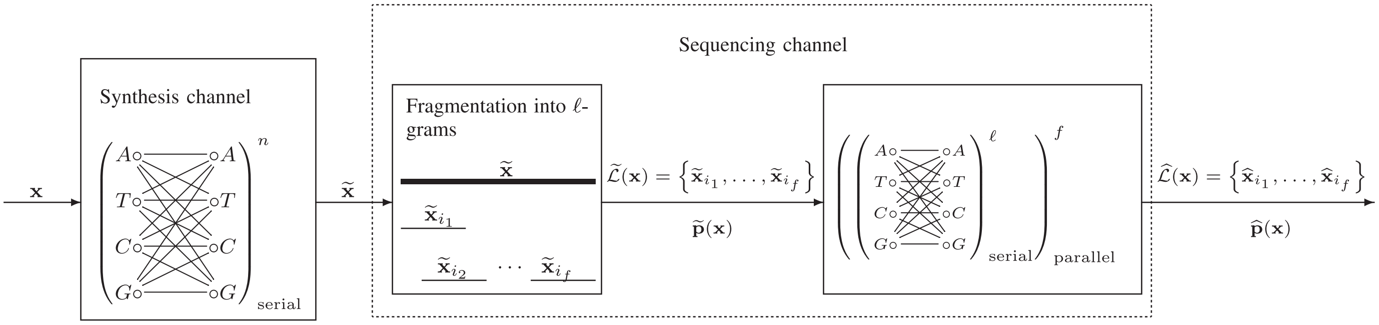


Fig. 1. The DNA Storage Channel. Information is encoded in a DNA sequence $\mathbf{x}$ which is synthesized with potential errors. The output of the synthesis process is $\widetilde{\mathbf{x}}$. During readout, the sequence $\widetilde{\mathbf{x}}$ is passed through the sequencing channel, which fragments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

choice of *representative* for X in Section 8 and henceforth refer to elements in $\mathcal{Q}(n;S)$ by their representative words.

Let $\mathbf{p}\mathcal{Q}(n;S)$ denote the set of profile vectors of words in $\mathcal{Q}(n;S)$. Then, $|\mathbf{p}\mathcal{Q}(n;S)| = |\mathcal{Q}(n;S)|$.

Furthermore, let $\mathcal{C} \subseteq \mathcal{Q}(n;S)$. If $d = \min\{d_{\text{gram}}(\mathbf{x},\mathbf{y};S) : \mathbf{x},\mathbf{y} \in \mathcal{C}, \mathbf{x} \neq \mathbf{y}\}$, then $\mathcal{C}$ is called an $(n,d;S)$-*ℓ-gram reconstruction code (GRC)*, or, $(n,d;S)$-GRC, for short. The following proposition demonstrates that an ℓ-gram reconstruction code is able to correct synthesis and sequencing errors provided that its ℓ-gram distance is sufficiently large. We observe that synthesis errors have effects that are ℓ times "stronger" since the error "propagates" through multiple ℓ-grams.

Example 3.1. Let $S = [\![2]\!]^2$ and $n = 4$. Then $\mathcal{Q}(n;S)$ comprises:

- two equivalence classes of size three, $\{0110, 1011, 1101\}$ and $\{0010, 0100, 1001)\}$, corresponding to the profile vectors, $(0,1,1,1)$ and $(1,1,1,0)$, respectively;
- ten equivalence classes of size one, $\{0000\}$, $\{0001\}$, $\{0011\}$, $\{0101\}$, $\{0111\}$, $\{1000\}$, $\{1010\}$, $\{1100\}$, $\{1110\}$ and $\{1111\}$.

Therefore, the profile vectors corresponding to words in $\mathcal{Q}(n;S)$ are given by the set

$$\begin{aligned}\mathbf{p}\mathcal{Q}(n;S) = \{&(0,0,0,3),(0,0,1,2),(0,1,0,2),(0,1,1,1),\\ &(0,1,2,0),(0,2,1,0),(1,0,1,1),(1,1,0,1),\\ &(1,1,1,0),(2,0,1,0),(2,1,0,0),(3,0,0,0)\},\end{aligned}$$

and $|\mathcal{Q}(n;S)| = |\mathbf{p}\mathcal{Q}(n;S)| = 12$.

Proposition 3.2. An $(n,d;S)$-GRC can correct s_{syn} substitution errors due to synthesis, s_{seq} substitution errors due to sequencing and t coverage errors provided that $d > 2s_{\text{syn}}\ell + 2s_{\text{seq}} + t$.

Proof: Consider an $(n,d;S)$-GRC $\mathcal{C}$ and the set $\mathbf{p}(\mathcal{C}) = \{\mathbf{p}(\mathbf{x};S) : \mathbf{x} \in \mathcal{C}\}$. By construction, $\mathbf{p}(\mathcal{C})$ is an (N,d)-AECC with $N = |S|$ that corrects all asymmetric errors of L_1-weight $\le 2s_{\text{syn}}\ell + 2s_{\text{seq}} + t$.

Suppose that, on the contrary, $\mathcal{C}$ cannot correct s_{syn} substitution errors due to synthesis, s_{seq} substitution errors due to sequencing and t coverage errors. Then, there exist two distinct codewords $\mathbf{x},\mathbf{x}' \in \mathcal{C}$ and error vectors $\mathbf{e}_{\text{syn},+}$, $\mathbf{e}_{\text{syn},-}$, $\mathbf{e}_{\text{seq},+}$, $\mathbf{e}_{\text{seq},-}$, $\mathbf{e}_t$, $\mathbf{e}'_{\text{syn},+}$, $\mathbf{e}'_{\text{syn},-}$, $\mathbf{e}'_{\text{seq},+}$, $\mathbf{e}'_{\text{seq},-}$, $\mathbf{e}'_t$, such that $\widehat{\mathbf{p}}(\mathbf{x}) = \widehat{\mathbf{p}}(\mathbf{x}')$, that is, such that

$$\begin{aligned}&\mathbf{p}(\mathbf{x};S) + \mathbf{e}_{\text{syn},+} - \mathbf{e}_{\text{syn},-} + \mathbf{e}_{\text{seq},+} - \mathbf{e}_{\text{seq},-} - \mathbf{e}_t\\ &\quad= \mathbf{p}(\mathbf{x}';S) + \mathbf{e}'_{\text{syn},+} - \mathbf{e}'_{\text{syn},-} + \mathbf{e}'_{\text{seq},+} - \mathbf{e}'_{\text{seq},-} - \mathbf{e}_t.\end{aligned}$$

Here, $\mathbf{e}_{\text{syn},-} - \mathbf{e}_{\text{syn},+}$ and $\mathbf{e}'_{\text{syn},-} - \mathbf{e}'_{\text{syn},+}$ are the error vectors due to substitutions during synthesis in $\mathbf{x}$ and $\mathbf{x}'$, respectively; each of the vectors $\mathbf{e}_{\text{syn},-}$, $\mathbf{e}_{\text{syn},+}$, $\mathbf{e}'_{\text{syn},-}$, $\mathbf{e}'_{\text{syn},+}$ has L_1-weight $s_{\text{syn}}\ell$; the vectors $\mathbf{e}_{\text{seq},-} - \mathbf{e}_{\text{seq},+}$ and $\mathbf{e}'_{\text{seq},-} - \mathbf{e}'_{\text{seq},+}$ model substitution errors during sequencing in $\mathbf{x}$ and $\mathbf{x}'$, respectively; each of the vectors $\mathbf{e}_{\text{seq},-}$, $\mathbf{e}_{\text{seq},+}$, $\mathbf{e}'_{\text{seq},-}$, $\mathbf{e}'_{\text{seq},+}$ has L_1-weight s_{seq}; and $\mathbf{e}_t$ and $\mathbf{e}'_t$ are the coverage error vectors of $\mathbf{x}$ and $\mathbf{x}'$, respectively, and both $\mathbf{e}_t, \mathbf{e}'_t$ have L_1-weight t. Therefore,

$$\begin{aligned}&\mathbf{p}(\mathbf{x};S) - (\mathbf{e}_{\text{syn},-} + \mathbf{e}_{\text{seq},-} + \mathbf{e}_t + \mathbf{e}'_{\text{syn},+} + \mathbf{e}'_{\text{seq},+})\\ &\quad= \mathbf{p}(\mathbf{x}';S) - (\mathbf{e}'_{\text{syn},-} + \mathbf{e}'_{\text{seq},-} + \mathbf{e}'_t + \mathbf{e}_{\text{syn},+} + \mathbf{e}_{\text{seq},+}),\end{aligned}$$

where $\mathbf{e}_{\text{syn},-} + \mathbf{e}_{\text{seq},-} + \mathbf{e}_t + \mathbf{e}'_{\text{syn},+} + \mathbf{e}'_{\text{seq},+}$ and $\mathbf{e}'_{\text{syn},-} + \mathbf{e}'_{\text{seq},-} + \mathbf{e}'_t + \mathbf{e}_{\text{syn},+} + \mathbf{e}_{\text{seq},+}$ are nonnegative vectors of L_1-weight at most $2s_{\text{syn}}\ell + 2s_{\text{seq}} + t$. This contradicts the fact that $\mathbf{p}(x;S)$ and $\mathbf{p}(\mathbf{x}';S)$ belong to a code that corrects asymmetric errors with L_1-weight at most $2s_{\text{syn}}\ell + 2s_{\text{seq}} + t$.

□

Throughout the remainder of the paper, we consider the problem of enumerating the profile vectors in $\mathbf{p}\mathcal{Q}(n;S)$ and constructing $(n,d;S)$-ℓ-gram reconstruction codes for a general subset $S \subseteq [\![q]\!]^\ell$. Our solutions are characterized by properties associated with a class of graphs defined on S, which we introduce in Section 4. In the same section, we collect enumeration results for $\mathcal{Q}(n;S)$. Section 5 is devoted to the proof of the main enumeration result using Ehrhart theory. We further exploit Ehrhart theory and certain graph theoretic concepts to construct codes in Section 6 and summarize numerical results for the special case where $S = S(q,\ell;q^*,[w_1,w_2])$ in Section 7. Finally, we describe practical decoding procedures in Section 8.

Remark 1.

For the case $S = [\![q]\!]^\ell$, given a word $\mathbf{x} \in [\![q]\!]^n$, Ukkonen observed certain properties of words belonging to the equivalence class of $\mathbf{x}$ [19]. Pevzner, based on Ukkonen's conjectures, then completely characterized all words within the equivalence class [19], [20]. In this paper, we focus on computing the *number* of equivalence classes for a general subset S, and to the best of our knowledge, this is the first work in this direction.

(ii) For ease of exposition, we abuse notation by identifying words in $\mathcal{Q}(n; S)$ with their corresponding profile vectors in $\mathbf{p}\mathcal{Q}(n; S)$ and refer to GRCs as being subsets of $\mathcal{Q}(n; S)$ or $\mathbf{p}\mathcal{Q}(n; S)$ interchangeably.

(iii) Given $(n, d; S)$-GRC $\mathcal{C}$ and the set $\mathbf{p}(\mathcal{C}) = \{\mathbf{p}(\mathbf{x}; S) : \mathbf{x} \in \mathcal{C}\}$, observe that all profile vectors in $\mathbf{p}(\mathcal{C})$ have L_1-weight $n - \ell + 1$. In this case, the asymmetric distance between two profile vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbf{p}(\mathcal{C})$ is given by half of the L_1-weight of $(\mathbf{u} - \mathbf{v})$. Therefore, with appropriate modifications, we may use codes constructed over L_1-distance to seed the constructions given in Section 6.

4. Restricted De Bruijn Graphs and Enumeration of Profile Vectors

We use standard concepts and terminology from graph theory, following Bollobás [21].

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 allow loops in our digraphs: in other words, we allow $v = v'$. In some instances, we allow multiple arcs between nodes and we term these digraphs as *multigraphs*.

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

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

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, it is known that the rank of $\mathbf{B}(D)$ equals $|V| - 1$ (see [21, §II, Thm 9 and Ex. 38]).

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 nodes, i.e., $v_i \neq v_j$, for $0 \leq i < j < n$. We consider a loop to be a cycle of length one. 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 $\mathbf{B}(D)\boldsymbol{\chi}(C) = \mathbf{0}$.

A closed walk is *Eulerian* if it includes all arcs in E. A cycle is *Hamiltonian* if it includes all nodes in V. A digraph is *strongly connected* if for all $v, v' \in V$, there exists a walk from v to v' and vice versa. A necessary and sufficient condition for a strongly connected graph to have a closed Eulerian walk is that the number of incoming arcs is equal to the number of outgoing arcs for each node. Furthermore, we have the following lemma. Here $\mathbf{u} > \mathbf{0}$ means that $\mathbf{u}_e > 0$ for each edge e.

Lemma 4.1. If D is strongly connected, then there exists a vector $\mathbf{u} > \mathbf{0}$ such that $\mathbf{B}(D)\mathbf{u} = \mathbf{0}$.

Proof: Let $D = (V, E)$. Since D is strongly connected, for each arc $vv' \in E$ there is a walk W from v' to v. Let $W_{vv'} = W \cup \{vv'\}$. Now $W_{vv'}$ is a closed walk containing the arc vv', so $\mathbf{B}(D)\boldsymbol{\chi}(W \cup \{vv'\}) = \mathbf{0}$. Therefore, the vector $\mathbf{u} = \sum_{vv' \in E} \boldsymbol{\chi}(C_{vv'})$ satisfies $\mathbf{B}(D)\mathbf{u} = \mathbf{0}$. It is easy to verify that $\mathbf{u} > \mathbf{0}$ since, for every arc vv', the entry corresponding to the arc vv' is at least 1. □

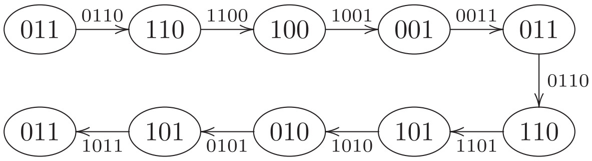
We are concerned with a special family of digraphs, namely, the de Bruijn graphs [22]. Given q and ℓ, the standard *de Bruijn graph* is defined on the node set $[\![q]\!]^{\ell-1}$. For $\mathbf{v}, \mathbf{v}' \in [\![q]\!]^{\ell-1}$, the ordered pair $(\mathbf{v}, \mathbf{v}')$ belongs to the arc set if and only if $v_i = v'_{i-1}$ for $2 \leq i \leq \ell - 1$. We label the arc $(\mathbf{v}, \mathbf{v}')$ with the length-ℓ word $\mathbf{v}v'_{\ell-1}$, and refer to arcs by these labels. (Note that $\mathbf{v}v'_{\ell-1} = v_1\mathbf{v}'$ if and only if $(\mathbf{v}, \mathbf{v}')$ is an arc.)

Example 4.1. Let $q = 2$, $\ell = 4$. Then the nodes $\mathbf{v} = 101$ and $\mathbf{v}' = 010$ are connected by the arc 1010 which originates from $\mathbf{v}$ and terminates in $\mathbf{v}'$ as the suffix of $\mathbf{v}$ of length $\ell - 2 = 2$ equals 01, which is also the prefix of length $\ell - 2$ of $\mathbf{v}'$.

The notion of restricted de Bruijn graphs was introduced by Ruskey *et al.* [23] for the case of a binary alphabet. For a fixed subset $S \subseteq [\![q]\!]^{\ell}$, we define the corresponding *restricted de Bruijn graph*, denoted by $D(S)$ as follows. The nodes of $D(S)$, denoted by $V(S)$, are the $(\ell - 1)$-grams appearing in the set 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 $v_1 v_2 \cdots v_{\ell-1} v'_{\ell-1} \in S$. Note that the standard de Bruijn graph is simply $D([\![q]\!]^{\ell})$. We refer the readers to Fig. 2 for an illustration of a de Bruijn and restricted de Bruijn graph with sets $[\![2]\!]^3$ and $S(2, 4; 1, [2, 3])$, respectively.

Example 4.2. Continuing Example 4.1, let $q = 2$, $\ell = 4$ and $S = S(2, 4; 1, [2, 3])$. Since the word 1010 belongs to S, the arc from $\mathbf{v} = 101$ and $\mathbf{v}' = 010$ belongs to $D(S)$. We also observe that 1010 is word of length $n = 4$ and it can be represented by the walk of length $n - \ell + 1 = 1$ from $\mathbf{v}$ to $\mathbf{v}'$.

In general, a word of length n whose ℓ-grams belong to S can be represented by a walk of length $n - \ell + 1$ in $D(S)$. For example, the word 011001101011 of length twelve corresponds to the walk



of length nine. Conversely, given the above *walk* of length nine, it is not difficult to obtain the binary word of length twelve. For each arc $\mathbf{z}$ in S, we observe that the number of times $\mathbf{z}$ is traversed by the walk gives the number of times of $\mathbf{z}$ appears as a 4-gram of the word. Hence, if we label each arc $\mathbf{z}$ by this number, we obtain a representation of the profile vector on $D(S)$. We refer the readers to Fig. 2 for an illustration.

In their paper, Ruskey *et al.* showed that $D(S)$ is Eulerian when $S = S(2, \ell; 1, [w - 1, w])$ for $w \in [\ell]$. Nevertheless, the results of [23] can be extended for general q, q^* and more general range of weights. As these extensions are needed for our subsequent derivation, we provide their technical proofs in Appendix B. For purposes of brevity, we write $D(S(q, \ell; q^*, [w_1, w_2]))$ and $D([\![q]\!]^{\ell})$ as

$D(q,\ell;q^*,[w_1,w_2])$ and $D(q,\ell)$, respectively.

Proposition 4.2. *Fix q and ℓ. Let $1 \leq q^* \leq q-1$ and $1 \leq w_1 < w_2 \leq \ell$. Then $D(q,\ell;q^*,[w_1,w_2])$ is Eulerian. In addition, $D(q,\ell)$ is Hamiltonian.*

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

We provide next 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 $D(S)$. Observe from Example 4.2 that a walk from node $\mathbf{v}$ to node $\mathbf{v}'$ in $D(S)$ is equivalent to a word whose ℓ-grams belong to S that starts with $\mathbf{v}$ and ends with $\mathbf{v}'$. Therefore, we define *closed words* to be words that start and end with the same $(\ell-1)$-gram to correspond with closed walks in $D(S)$. 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 $\mathbf{p}\bar{\mathcal{Q}}(n;S)$. Clearly, $\bar{\mathcal{Q}}(n;S) \subseteq \mathcal{Q}(n;S)$ and as illustrated in the next example, $\bar{\mathcal{Q}}(n;S)$ is properly contained in $\mathcal{Q}(n;S)$ for most cases.

Example 4.3. Consider again the setting where $S = [\![2]\!]^2$ and $n=4$. The set of closed words is

$$\{0000, 0010, 0100, 0110, 1001, 1011, 1101, 1111\},$$

while the equivalence classes in $\bar{\mathcal{Q}}(n;S)$ equal

$$\{0000\}, \{0010, 0100, 1001\}, \{0110, 1011, 1101\}, \{1111\}.$$

The profile vectors corresponding to words in $\mathbf{p}\bar{\mathcal{Q}}(n;S)$ are $(3,0,0,0)$, $(1,1,1,0)$, $(0,1,1,1)$ and $(0,0,0,3)$.

We observe that words not in $\bar{\mathcal{Q}}(n;S)$, such as 0111, have profile vectors that do not belong to $\mathbf{p}\bar{\mathcal{Q}}(n;S)$.

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:

$$\mathbf{B}(D(S))\mathbf{u} = \mathbf{0}. \tag{1}$$

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

$$\mathbf{1}^T\mathbf{u} = n-\ell+1. \tag{2}$$

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. Equations (1) and (2) may then be rewritten as $\mathbf{A}(S)\mathbf{u} = (n-\ell+1)\mathbf{b}$.

Consider the following two sets of integer points

$$\mathcal{LP}_{\geq 0}(n;S) \triangleq \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}(S)\mathbf{u} = (n-\ell+1)\mathbf{b},\ \mathbf{u} \geq \mathbf{0}\}, \tag{3}$$

$$\mathcal{LP}_{>0}(n;S) \triangleq \{\mathbf{u} \in \mathbb{Z}^{|S|} : \mathbf{A}(S)\mathbf{u} = (n-\ell+1)\mathbf{b},\ \mathbf{u} > \mathbf{0}\}. \tag{4}$$

The preceding discussion asserts that the profile vector of any closed word must lie in $\mathcal{LP}_{\geq 0}(n;S)$. Conversely, the next lemma shows that any vector in $\mathcal{LP}_{>0}(n;S)$ is a profile vector of some word in $\bar{\mathcal{Q}}(n;S)$.

Lemma 4.3. *Suppose that $D(S)$ is strongly connected. If $\mathbf{u} \in \mathcal{LP}_{>0}(n;S)$, then there exists a word $\mathbf{x} \in \bar{\mathcal{Q}}(n;S)$ such that $\mathbf{p}(\mathbf{x};S) = \mathbf{u}$. That is, $\mathcal{LP}_{>0}(n;S) \subseteq \mathbf{p}\bar{\mathcal{Q}}(n;S)$.*

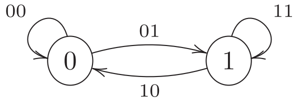
Proof: Construct a multidigraph $D_\mathbf{u}$ on the node set $V(S)$ such that there are $u_\mathbf{z}$ copies of the arc $\mathbf{z}$ for all $\mathbf{z} \in S$. Since each $u_\mathbf{z}$ is positive and $D(S)$ is strongly connected, $D_\mathbf{u}$ is also strongly connected. Since $\mathbf{u} \in \mathcal{LP}_{>0}(n;S)$, $\mathbf{u}$ also satisfies the flow conservation equations and $D_\mathbf{u}$ is consequently Eulerian. Also, as $D_\mathbf{u}$ has $n-\ell+1$ arcs, an Eulerian walk on $D_\mathbf{u}$ yields one such desired word $\mathbf{x}$. □

Therefore, we have the following relation:

$$\mathcal{LP}_{>0}(n;S) \subseteq \mathbf{p}\bar{\mathcal{Q}}(n;S) \subseteq \mathcal{LP}_{\geq 0}(n;S). \tag{5}$$

Example 4.4. We illustrate next through two examples how given S one may determine the sets $\mathbf{p}\mathcal{Q}(n;S)$, $\mathbf{p}\bar{\mathcal{Q}}(n;S)$, $\mathcal{LP}_{>0}(n;S)$, and $\mathcal{LP}_{\geq 0}(n;S)$, as well as enumerate $\mathcal{LP}_{\geq 0}(n;S)$.

(a) Let $S = [\![2]\!]^2$. Then, $D(S)$ is given by



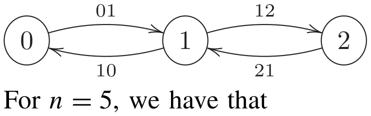
For $n=5$, we have

$$\begin{aligned}\mathbf{p}\mathcal{Q}(5;S) = \{&(0,0,0,4), (0,0,1,3), (0,1,0,3),\\ &(0,1,1,2), (0,1,2,1), (0,2,1,1),\\ &(0,2,2,0), (1,0,1,2), (1,1,0,2),\\ &(1,1,1,1), (1,1,2,0), (1,2,1,0),\\ &(2,0,1,1), (2,1,0,1), (2,1,1,0),\\ &(3,0,1,0), (3,1,0,0), (4,0,0,0)\},\\ \mathbf{p}\bar{\mathcal{Q}}(5;S) = \{&(0,0,0,4), (0,1,1,2), (0,2,2,0),\\ &(1,1,1,1), (2,1,1,0), (4,0,0,0)\}, \text{ and}\\ \mathcal{LP}_{\geq 0}(5;S) = \{&(0,0,0,4), (0,1,1,2), (0,2,2,0),\\ &(1,0,0,3), (1,1,1,1), (2,0,0,2),\\ &(2,1,1,0), (3,0,0,1), (4,0,0,0)\},\\ \mathcal{LP}_{>0}(5;S) = \{&(1,1,1,1)\}.\end{aligned}$$

It is straightforward to verify that, indeed, $\mathcal{LP}_{>0}(5;S) \subseteq \mathbf{p}\bar{\mathcal{Q}}(5;S) \subseteq \mathcal{LP}_{\geq 0}(5;S)$. Furthermore, a simple calculation shows that

$$\mathcal{LP}_{\geq 0}(n;[\![2]\!]^2) = \begin{cases} \frac{n^2}{4} + \frac{n}{2}, & \text{if } n \text{ is even,} \\ \frac{n^2}{4} + \frac{n}{2} + \frac{1}{4}, & \text{otherwise.} \end{cases}$$

(b) Let $q=3$, $\ell=2$, and $S = \{01, 10, 12, 21\}$. Then $D(S)$ is given by



For $n=5$, we have that

$$\begin{aligned}\mathbf{p}\mathcal{Q}(5;S) = \{&(0,0,2,2), (0,1,1,2), (1,0,2,1),\\ &(1,1,1,1),\\ &(1,2,0,1), (2,1,1,0), (2,2,0,0)\}\\ \mathbf{p}\bar{\mathcal{Q}}(5;S) = \{&(0,0,2,2), (1,1,1,1), (2,2,0,0)\}, \text{ and}\\ \mathcal{LP}_{\geq 0}(5;S) = \{&(0,0,2,2), (1,1,1,1), (2,2,0,0)\},\\ \mathcal{LP}_{>0}(5;S) = \{&(1,1,1,1)\}.\end{aligned}$$

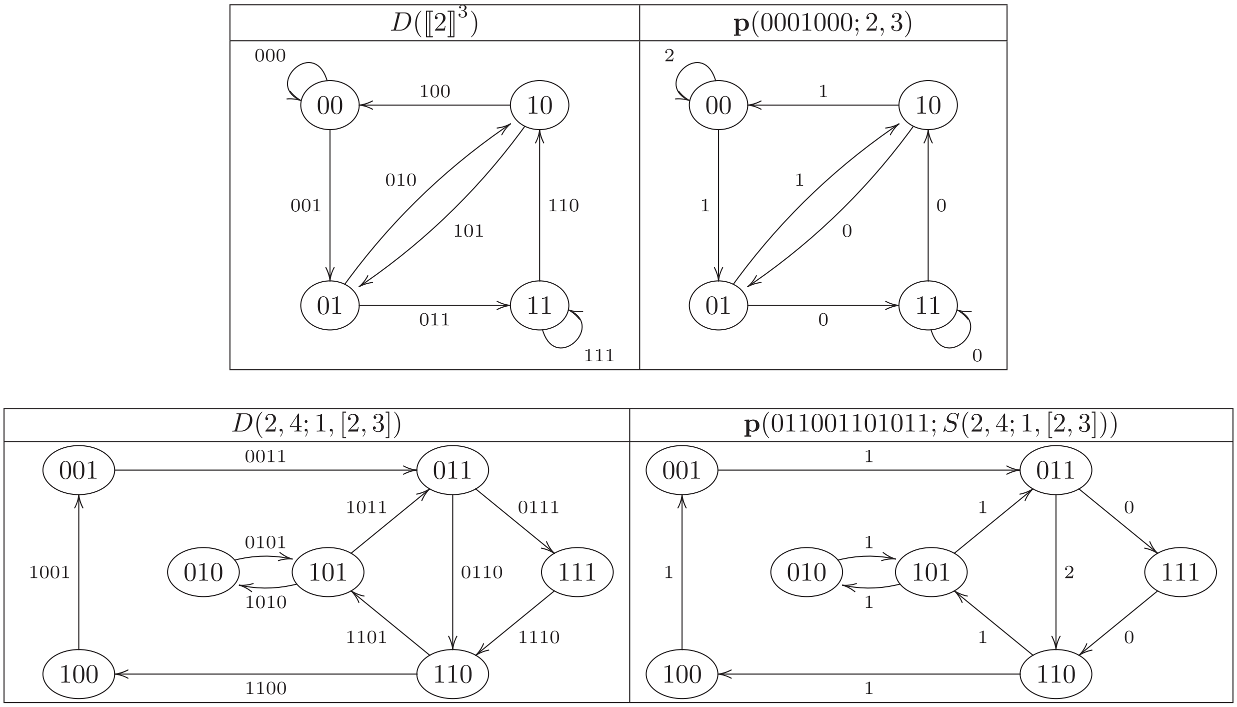


Fig. 2. Examples of two de Bruijn and restricted de Bruijn graphs. The upper left corner shows a classical de Bruijn graph with $q = 2$ and $\ell = 3$. Note that the nodes of the graph are all binary tuples of length $\ell - 1 = 2$, and arcs in the graph connect any pair of nodes for which the last symbol of the origin node equals the first symbol of the terminal node. The arcs are labeled by the "overlap" sequence of the node labels. In the right hand corner, the same graph is depicted with respect to a input sequence $\mathbf{x}$ which induces weights on the arcs, indicating how many times the ℓ-gram corresponding to the arc appeared in $\mathbf{x}$. For example, in $\mathbf{x} = 0001000$, the $\ell = 3$-gram appears twice, leading to the label 2 for the self-loop around the node 00. This example is extended for the case of a restricted de Bruijn graph defined on the set $S(2, 4; 1, [2, 3])$ as depicted in the second row. Note that the graph in the lower left corner contains only arcs labeled by $\ell = 4$-tuples of weight 2 and 3, as required by the definition of $S(2, 4; 1, [2, 3])$. The corresponding 4-gram profile vector for 011001101011 on the aforementioned restricted de Bruijn graph is shown in the lower right corner. As an example, observe that the sequence $\mathbf{x} = 011001101011$ has two substrings 0110, and hence the arc from the node labeled by 011 to the node labeled by 110 has weight 2.

We again verify that $\mathcal{LP}_{>0}(5; S) \subseteq \mathbf{p}\bar{\mathcal{Q}}(5; S) \subseteq \mathcal{LP}_{\geq 0}(5; S)$. In addition,

$$\mathcal{LP}_{\geq 0}(n; \{01, 10, 12, 21\}) = \begin{cases} 0, & \text{if } n \text{ is even,} \\ \frac{n}{2} + \frac{1}{2}, & \text{otherwise.} \end{cases}$$

We first state our main enumeration result and defer its proof to Section 5. Specifically, under the assumption that $D(S)$ is strongly connected, we show that both $|\mathcal{LP}_{>0}(n; S)|$ and $|\mathcal{LP}_{\geq 0}(n; S)|$ are quasipolynomials in n whose coefficients are periodic in n. Following Beck and Robins [24], we define a *quasipolynomial* f as a function in n of the form $c_D(n)n^D + c_{D-1}(n)n^{D-1} + \cdots + c_0(n)$, where $c_D, c_{D-1}, \ldots, c_0$ are periodic functions in n. 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 $c_D, c_{D-1}, \ldots, c_0$.

Example 4.5. From above, we see that $\mathcal{LP}_{\geq 0}(n; [\![2]\!]^2)$ and $\mathcal{LP}_{\geq 0}(n; \{01, 10, 12, 21\})$ are quasipolynomials of degrees two and one, respectively. The periods of both quasipolynomials are two. We note that even though $\mathcal{LP}_{\geq 0}(n; \{01, 10, 12, 21\})$ is a quasipolynomial of degree one, $\mathcal{LP}_{\geq 0}(n; \{01, 10, 12, 21\}) \neq \Theta(n)$ as the function evaluates to zero when n is even. Hence, we adapt the usual Θ-notation to capture the periodic behaviour of $\mathcal{LP}_{\geq 0}(n; S)$.

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 $f(n) \geq cg(n)$ for sufficiently large n with $\lambda | (n - \ell + 1)$. In other words, $f(n) \geq cg(n)$ whenever n is sufficiently large and is congruent to $\ell - 1$ modulo λ. We write $f(n) = \Theta'(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega'(g(n))$.

Theorem 4.4. Suppose $D(S)$ is strongly connected and let λ be the least common multiple of the lengths of all cycles in $D(S)$. Then $|\mathcal{LP}_{>0}(n; S)| = \Theta'\left(n^{|S|-|V(S)|}\right)$ and $|\mathcal{LP}_{\geq 0}(n; S)| = \Theta'\left(n^{|S|-|V(S)|}\right)$. In particular, $|\mathbf{p}\bar{\mathcal{Q}}(n; S)| = \Theta'\left(n^{|S|-|V(S)|}\right)$.

As illustrated by Examples 4.4 and 4.5, the size of $\mathbf{p}\bar{\mathcal{Q}}(n; S)$ can possibly evaluate to zero for infinitely many values of n. However, Theorem 4.4 guarantees the existence of some period λ, whereby the size of $\mathbf{p}\bar{\mathcal{Q}}(n; S)$ is proportional to $n^{|S|-|V(S)|}$ whenever n is congruent to $\ell - 1$ modulo λ.

Before we end this section, we look at certain implications of Theorem 4.4. First, we show that the estimate on $|\mathbf{p}\bar{\mathcal{Q}}(n; S)|$ extends to $|\mathbf{p}\mathcal{Q}(n; S)|$ when $D(S)$ is strongly connected.

Corollary 4.5. 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

$$|\mathbf{p}\mathcal{Q}(n; S)| = \Theta'(|\mathbf{p}\mathcal{Q}(n; S, \mathbf{z} \to \mathbf{z}')|)$$

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

Proof: Let $\mathbf{z}, \mathbf{z}' \in V(S)$. Since $D(S)$ is strongly connected, we consider the shortest path from $\mathbf{z}$ to $\mathbf{z}'$ in $D(S)$. Let $\mathbf{w} = \mathbf{z}\mathbf{w}'$ be the corresponding q-ary word and $L(\mathbf{z}, \mathbf{z}')$ be the length of the path, or equivalently, the length of the word $\mathbf{w}'$. Consider $\mathbf{u}(\mathbf{z} \to \mathbf{z}') = \mathbf{p}(\mathbf{w}; S)$ the profile vector of $\mathbf{w}$ and observe that both the length $L(\mathbf{z}, \mathbf{z}')$ and the vector $\mathbf{u}(\mathbf{z} \to \mathbf{z}')$ are independent of n.

We demonstrate the following inequality:

$$\begin{aligned} |\mathcal{LP}_{>0}(n - L(\mathbf{z}, \mathbf{z}'); S)| &\le |\mathbf{p}\mathcal{Q}(n; S, \mathbf{z} \to \mathbf{z}')| \\ &\le |\mathbf{p}\bar{\mathcal{Q}}(n + L(\mathbf{z}', \mathbf{z}); S)|. \end{aligned} \quad (6)$$

To demonstrate the first inequality, we construct an injective map $\phi_1 : \mathcal{LP}_{>0}(n - L(\mathbf{z}, \mathbf{z}'); S) \to \mathbf{p}\mathcal{Q}(n; S, \mathbf{z} \to \mathbf{z}')$ defined by $\phi_1(\mathbf{u}) = \mathbf{u} + \mathbf{u}(\mathbf{z} \to \mathbf{z}')$. Now, since $\mathbf{u} \in \mathcal{LP}_{>0}(n - L(\mathbf{z}, \mathbf{z}'); S)$, we obtain from Lemma 4.3 a word of length $n - L(\mathbf{z}, \mathbf{z}')$ whose profile vector is $\mathbf{u}$. Without loss of generality, we let this word be $\mathbf{x}$ and assume that it starts and ends with $\mathbf{z}$. Then $\mathbf{x}\mathbf{w}'$ is a word of length n whose profile vector is $\mathbf{u}+\mathbf{u}(\mathbf{z} \to \mathbf{z}')$. Therefore, $\phi_1(\mathbf{u})$ lies in $\mathbf{p}\mathcal{Q}(n; S, \mathbf{z} \to \mathbf{z}')$ and ϕ_1 is a well-defined map. Suppose $\mathbf{u}$ and $\mathbf{u}'$ are vectors in $\mathcal{LP}_{>0}(n - L(\mathbf{z}, \mathbf{z}'); S)$ such that $\phi_1(\mathbf{u}) = \phi_1(\mathbf{u}')$. Since $\mathbf{u} = \phi_1(\mathbf{u}) - \mathbf{u}(\mathbf{z} \to \mathbf{z}') = \phi_1(\mathbf{u}') - \mathbf{u}(\mathbf{z} \to \mathbf{z}') = \mathbf{u}'$, we conclude ϕ_1 is injective and hence, the first inequality follows.

Similarly, for the other inequality, we consider another map $\phi_2 : \mathbf{p}\mathcal{Q}(n; S, \mathbf{z} \to \mathbf{z}') \to \mathbf{p}\bar{\mathcal{Q}}(n + L(\mathbf{z}', \mathbf{z}); S)$ where $\phi_2(\mathbf{u}) = \mathbf{u}+\mathbf{u}(\mathbf{z}' \to \mathbf{z})$. As before, let $\mathbf{u}$ be the profile vector of a word $\mathbf{x}$ of length n that starts with $\mathbf{z}$ and ends with $\mathbf{z}'$. Let $\mathbf{w} = \mathbf{z}'\mathbf{w}'$ be the q-ary word corresponding to the shortest path from $\mathbf{z}'$ to $\mathbf{z}$ in $D(S)$. Concatenating $\mathbf{x}$ with $\mathbf{w}'$ yields $\mathbf{x}\mathbf{w}'$, which is a word of length $n + L(\mathbf{z}', \mathbf{z})$ and starts and ends with $\mathbf{z}$. Hence, its profile vector $\mathbf{u} + \mathbf{u}(\mathbf{z}' \to \mathbf{z})$ lies in $\mathbf{p}\bar{\mathcal{Q}}(n + L(\mathbf{z}', \mathbf{z}); S)$. As with ϕ_1, the map ϕ_2 is well-defined and can be shown to be injective.

Combining (6) with Theorem 4.4 yields the result $|\mathbf{p}\mathcal{Q}(n; S, \mathbf{z}, \mathbf{z}')| = \Theta'\left(n^{|S|-|V(S)|}\right)$.

Next, we demonstrate that $|\mathbf{p}\mathcal{Q}(n; S)| = \Theta'\left(n^{|S|-|V(S)|}\right)$, and observe that the other asymptotic equalities may be derived similarly.

Let $P \triangleq \max\{L(\mathbf{z}, \mathbf{z}') : \mathbf{z}, \mathbf{z}' \in V(S)\}$ be the diameter of the digraph $D(S)$. Then,

$$\begin{aligned} |\mathbf{p}\mathcal{Q}(n; S)| &= \sum_{\mathbf{z}, \mathbf{z}' \in V(S)} |\mathcal{Q}(n; S, \mathbf{z}, \mathbf{z}')| \\ &\le \sum_{\mathbf{z}, \mathbf{z}' \in V(S)} |\bar{\mathcal{Q}}(n + L(\mathbf{z}', \mathbf{z}); S)| \\ &\le |V(S)|^2 |\bar{\mathcal{Q}}(n + P; S)| = O\left(n^{|S|-|V(S)|}\right). \end{aligned}$$

Since $|\mathcal{Q}(n; S)| \ge |\bar{\mathcal{Q}}(n; S)| = \Omega'\left(n^{|S|-|V(S)|}\right)$, the corollary follows. □

In the special case where $S = \llbracket q \rrbracket^\ell$, Jacquet *et al.* demonstrated a stronger version of Theorem 4.4 for the special case $\ell = 2$ using analytic combinatorics. In addition, using a careful analysis similar to the proof of Corollary 4.5, Jacquet *et al.* also provided a tighter bound for $|\mathbf{p}\mathcal{Q}(n; \llbracket q \rrbracket^\ell)|$ for the case $\ell = 2$. Note that $f(n) \sim g(n)$ stands for $\lim_{n\to\infty} f(n)/g(n) = 1$.

Theorem 4.6 (Jacquet *et al.* [15])**.** Fix q, ℓ. Let $\mathcal{LP}_{>0}(n; \llbracket q \rrbracket^\ell)$, $\mathcal{LP}_{\ge 0}(n; \llbracket q \rrbracket^\ell)$, $\mathbf{p}\mathcal{Q}(n; \llbracket q \rrbracket^\ell)$ and $\mathbf{p}\bar{\mathcal{Q}}(n; \llbracket q \rrbracket^\ell)$ be defined as above. Then

$$\begin{aligned} |\mathcal{LP}_{>0}(n; \llbracket q \rrbracket^\ell)| &\sim |\mathcal{LP}_{\ge 0}(n; \llbracket q \rrbracket^\ell)| \\ &\sim |\mathbf{p}\bar{\mathcal{Q}}(n; \llbracket q \rrbracket^\ell)| \sim c(q, \ell) n^{q^\ell - q^{\ell-1}}, \end{aligned} \quad (7)$$

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

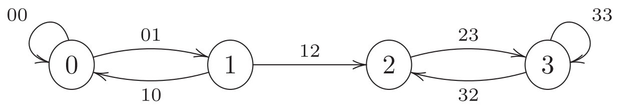
Next, we extend Theorem 4.4 to provide estimates on $\bar{\mathcal{Q}}(n; S)$ and $\mathcal{Q}(n; S)$ for general S, where $D(S)$ is not necessarily strongly connected.

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 $1 \le i \le I$. Define $\delta_i \triangleq |S_i| - |V_i|$. By Theorem 4.4, if S_i is nonempty then there are $\Theta'(n^{\delta_i})$ closed words belonging to $\bar{\mathcal{Q}}(n; S_i)$; these words also belong to $\bar{\mathcal{Q}}(n; S)$. On the other hand, if $S_i = \emptyset$ then clearly $\bar{\mathcal{Q}}(n; S_i) = \emptyset$ as well. Suppose $\bar{\Delta} = \max\{\delta_i : 1 \le i \le I\}$. Then $|\bar{\mathcal{Q}}(n; S)| = \Omega'(n^{\bar{\Delta}})$ unless $\bar{\Delta} = -1$, in which case $D(S)$ is acyclic so that $|\bar{\mathcal{Q}}(n; S)| = 0$.

On the other hand, any closed word $\mathbf{x}$ in $\bar{\mathcal{Q}}(n; S)$ corresponds to a closed walk in $D(S)$ and a closed walk in $D(S)$ must belong to some strongly connected component (V_i, S_i). In other words, $\mathbf{x}$ must belong to $\bar{\mathcal{Q}}(n; S_i)$ for some $1 \le i \le I$. Hence, we have $|\bar{\mathcal{Q}}(n; S)| = O(n^{\bar{\Delta}})$.

Corollary 4.7. 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 strongly connected for all $1 \le i \le I$. Define $\bar{\Delta} \triangleq \max\{|S_i| - |V_i| : 1 \le i \le I\}$. If $\bar{\Delta} \ge 0$, then $|\bar{\mathcal{Q}}(n; S)| = \Theta'(n^{\bar{\Delta}})$. If $\bar{\Delta} = -1$, then $|\bar{\mathcal{Q}}(n; S)| = 0$.

Example 4.6. 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 4.7.

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 $\mathbf{u} = \mathbf{u}_1 + \chi(12) + \mathbf{u}_2$ 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 $|\mathbf{p}\bar{\mathcal{Q}}(n; S_1, * \to 1)| = |\mathbf{p}\bar{\mathcal{Q}}(n; S_2, 2 \to *)| = n - 1$ for $n \ge 2$. Hence,

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

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 for $1 \le i, j \le I$, we add an arc from v_i to v_j. Further, we add an arc from v_{source} to v_i and from v_i to v_{sink} for all $1 \le i \le I$. The arcs leaving v_{source} have zero weight. For all $1 \le i \le I$, 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. 3 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 (see Ahuja *et al.* [25, Ch. 4]). Furthermore, 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 4.8. 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 strongly connected for all $1 \le i \le I$. Construct D' as above (see Fig. 3) and let Δ be the weight of the longest weighted path from v_{source} to v_{sink}. Then $|\mathbf{p}\mathcal{Q}(n;S)| = \Theta'(n^{\Delta})$.

Proof: Let $K \subset \{1, \ldots, I\}$ be the set of indices k such that $S_k = \emptyset$. In other words, the induced subgraph (V_k, S_k) is an isolated node. Define ϵ_j to be 0 if $j \in K$ and δ_j otherwise.

For each $\mathbf{u} \in \mathbf{p}\mathcal{Q}(n;S)$, we have a set of indices $\{i_1, i_2, \ldots, i_t\} \subseteq \{1, 2, \ldots, I\}$, a set of vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_t$, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{t-1}$, and integers $n_1, n_2, \ldots, n_t$ such that the following hold:

(i) $\mathbf{u} = \mathbf{u}_1 + \mathbf{e}_1 + \mathbf{u}_2 + \mathbf{e}_2 + \cdots + \mathbf{e}_{t-1} + \mathbf{u}_t$;
(ii) for $1 \le j \le t-1$, $\mathbf{e}_i$ is the incidence vector of some arc $(\mathbf{z}_j, \mathbf{z}'_{j+1})$ in $D(S)$ such that $\mathbf{z}_j \in V_{i_j}$ and $\mathbf{z}'_{j+1} \in V_{i_{j+1}}$;
(iii) for $1 \le j \le t$, the vector $\mathbf{u}_j$ belongs[3] to $\mathbf{p}\mathcal{Q}(n_j + \ell - 1; S_{i_j})$;
(iv) $(t-1) + \sum_{j=1}^{t} n_j = n - \ell + 1$;
(v) $v_{\text{source}} v_{i_1} v_{i_2} \cdots v_{i_t} v_{\text{sink}}$ is a path in D'.

Note that Condition (iii) implies that $n_j = 0$ whenever $i_j \in K$. Note that if $\mathbf{u}, \mathbf{u}'$ are vectors in $\mathbf{p}\mathcal{Q}(n;S)$ having the same set of indices $\{i_1, \ldots, i_t\}$ and the same vectors $\mathbf{u}_1, \ldots, \mathbf{u}_t$, then $\mathbf{u} = \mathbf{u}'$. Thus, we may obtain an upper bound on $|\mathbf{p}\mathcal{Q}(n;s)|$ by bounding the number of ways to produce such index sets and vectors.

For a fixed subset $\{i_1, i_2, \ldots, i_t\} \subseteq \{1, 2, \ldots, I\}$, let $k = |\{i_1, \ldots, i_t\} \cap K|$. Let T be the set of nonnegative integer tuples $(n_1, \ldots, n_t)$ such that $\sum_{j=1}^{t} n_j = (n - \ell + 1) - (t-1)$ and such that $n_j = 0$ whenever $i_j \in K$. If $k < t$, then $|T| \le n^{t-1-k}$, so we have

$$\begin{aligned}
\sum_{(n_1,\ldots,n_t)\in T} \prod_{j=1}^{t} \left|\mathbf{p}\mathcal{Q}(n_j + \ell - 1; S_{i_j})\right| \\
= |T|\, O(n^{\epsilon_{i_1} + \cdots + \epsilon_{i_t}}) \\
= O(n^{t-1-k}) O(n^{\delta_{i_1} + \cdots + \delta_{i_t} + k}) \\
= O(n^{\delta_{i_1} + \cdots + \delta_{i_t} + (t-1)}) = O(n^{\Delta}).
\end{aligned}$$

Here, the first inequality follows from Corollary 4.5, while the last inequality follows from the fact that $(t-1) + \sum_{j=1}^{t} \delta_{i_j}$ measures the weight of $v_{\text{source}} v_{i_1} v_{i_2} \cdots v_{i_t} v_{\text{sink}}$ and this value is upper bounded by Δ. On the other hand, if $k = t$, that is, if $\{i_1, \ldots, i_t\} \subseteq K$, then $|T| = 0$ if $t - 1 < n - \ell + 1$ and $|T| = 1$ otherwise. Hence in this case we also have $\sum_{(n_1,\ldots,n_t)\in T} \prod_{j=1}^{t} \left|\mathbf{p}\mathcal{Q}(n_j; S_{i_j})\right| = O(n^{\Delta})$. Since the number of subsets of $\{1, 2, \ldots I\}$ is independent of n, and since each subset corresponds to at most $O(n^{\Delta})$ vectors in $\mathbf{p}\mathcal{Q}(n;S)$, we have $|\mathbf{p}\mathcal{Q}(n;S)| = O(n^{\Delta})$.

Conversely, suppose $v_{\text{source}} v_{i_1} v_{i_2} \cdots v_{i_t} v_{\text{sink}}$ is a path in D' of maximum weight Δ. With T defined as before relative to $\{i_1, \ldots, i_t\}$, we then have

$$\begin{aligned}
|\mathbf{p}\mathcal{Q}(n;S)| &\ge \sum_{(n_1,\ldots,n_t)\in T} \prod_{j=1}^{t} |\mathbf{p}\mathcal{Q}(n_j + \ell - 1; S_{i_j})| \\
&\ge C_1 \sum_{(n_1,\ldots,n_t)\in T} n_1^{\epsilon_{i_1}} n_2^{\epsilon_{i_1}} \cdots n_t^{\epsilon_{i_t}}
\end{aligned}$$

for some positive constant C_1, by Corollary 4.5. (Note that we have adopted the convention that $0^0 = 1$.) Let $k = |K \cap \{i_1, \ldots, i_t\}|$ as before, and let $T' \subset T$ be the set defined by

$$T' = \left\{ (n_1, \ldots, n_t) \in T : n_j \ge \frac{n}{2t} \text{ whenever } i_j \notin K \right\}.$$

Observe that there is a positive constant C_2 such that for n sufficiently large, $|T'| \ge C_2 n^{(t-1)-k}$. Now we have

$$\begin{aligned}
\sum_{(n_1,\ldots,n_t)\in T'} n_1^{\epsilon_{i_1}} \cdots n_t^{\epsilon_{i_t}} &\ge (2t)^{-t} \sum_{(n_1,\ldots,n_t)\in T'} n^{\epsilon_{i_1} + \cdots + \epsilon_{i_t}} \\
&\ge (2t)^{-t} C_2 n^{\delta_{i_1} + \cdots + \delta_{i_t} + (t-1)} \\
&= C_3 n^{\Delta}.
\end{aligned}$$

□

5. Ehrhart Theory and Proof of Theorem 4.4

We assume $D(S)$ to be strongly connected and provide a detailed proof of Theorem 4.4. For this purpose, in the next subsection, we introduce some fundamental results from Ehrhart theory. Ehrhart theory is a natural framework for enumerating profile vectors and one may simplify the techniques of [15] significantly and obtain similar results for a more general family of digraphs. Furthermore, Ehrhart theory also allows us to extend the enumeration procedure to profiles at a prescribed distance.

A. Ehrhart Theory

As suggested by (3) and (4), in order to enumerate codewords of interest, we need to enumerate certain sets of integer points or lattice points in polytopes. The first general treatment of the theory of enumerating lattice points in polytopes was described by Ehrhart [26], and later developed by Stanley from a commutative-algebraic point of view (see [27, Ch. 4]). Here, we follow the combinatorial treatment of Beck and

[3] For ease of notation, we regard vectors in $\mathbf{p}\mathcal{Q}(n_j + \ell - 1; S_{i_j})$ as vectors in $\mathbf{p}\mathcal{Q}(n_j + \ell - 1; S)$ since $S_{i_j} \subseteq S$. If $\mathbf{u}$ belongs $\mathbf{p}\mathcal{Q}(n_j + \ell - 1; S_{i_j})$, we consider $\mathbf{u}' \in \mathbf{p}\mathcal{Q}(n_j + \ell - 1; S)$ where $\mathbf{u}'_{\mathbf{z}} = \mathbf{u}'_{\mathbf{z}}$ if $\mathbf{z} \in S_{i_j}$, and $\mathbf{u}'_{\mathbf{z}} = 0$, otherwise.

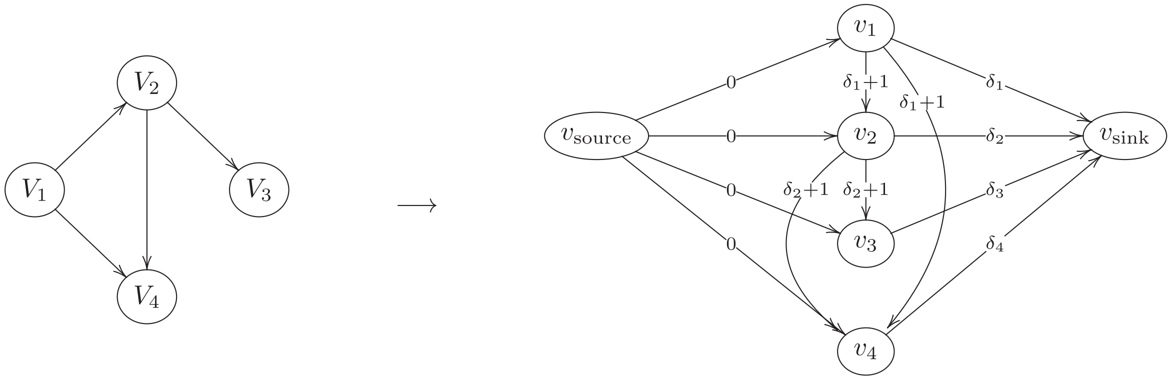


Fig. 3. Constructing a weighted digraph from the connected components of $D(S)$.

Robins [24]. Recall that $\mathbf{v} \geq \mathbf{0}$ means that all entries in $\mathbf{v}$ are nonnegative. We extend the notation so that $\mathbf{v} \geq \mathbf{u}$ denotes $\mathbf{v} - \mathbf{u} \geq \mathbf{0}$.

Consider the set $\mathcal{P}$ of points 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 $\mathbf{b}$. We then call this set $\mathcal{P}$ a *rational polytope*. A rational polytope is *integer* if all of its vertices (see Definition 5.1) have integer coordinates. For all positive integers t, let $t\mathcal{P}$ be the set $\{t\mathbf{u} : \mathbf{u} \in \mathcal{P}\}$. The *lattice point enumerator* $L_{\mathcal{P}}(t)$ of $\mathcal{P}$ is given by

$$L_{\mathcal{P}}(t) \triangleq |\mathbb{Z}^n \cap t\mathcal{P}|, \text{ for all postive integers } t.$$

Ehrhart [26] 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 5.1 (Ehrhart's theorem for polytopes [24, 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 (4), we consider the *relative interior* of $\mathcal{P}$. For the case where $\mathcal{P}$ 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 $L_{\mathcal{P}^{\circ}}(t)$, proved by Macdonald [28].

Theorem 5.2 (Ehrhart-Macdonald reciprocity [24, 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

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

B. Proof of Theorem 4.4

Recall the definitions of $\mathbf{A}(S)$ and $\mathbf{b}$ in (3), and consider the polytope

$$\mathcal{P}(S) \triangleq \{\mathbf{u} \in \mathbb{R}^{|S|} : \mathbf{A}(S)\mathbf{u} = \mathbf{b}, \mathbf{u} \geq \mathbf{0}\}, \tag{8}$$

Using lattice point enumerators, we may write $|\mathcal{LP}_{\geq 0}(n; S)| = L_{\mathcal{P}(S)}(n - \ell + 1)$. Therefore, in view of Ehrhart's theorem, we need to determine the dimension of the polytope $\mathcal{P}(S)$ and characterize the interior and the vertices of this polytope.

Lemma 5.3. Suppose that $D(S)$ is strongly connected. Then the dimension of $\mathcal{P}(S)$ is $|S| - |V(S)|$.

Proof: We first establish that the rank of $\mathbf{A}(S)$ is $|V(S)|$. Since $D(S)$ is connected, the rank of $\mathbf{B}(D(S))$ is $|V(S)| - 1$. We next show that $\mathbf{1}^T$ does not belong to the row space of $\mathbf{B}(D(S))$. As $D(S)$ is strongly connected, $D(S)$ contains a cycle, say C. Since $\mathbf{B}(D(S))\boldsymbol{\chi}(C) = 0$ but $\mathbf{1}\boldsymbol{\chi}(C) = |C| \neq 0$, $\mathbf{1}$ does not belong to the row space of $\mathbf{B}(D(S))$, so augmenting the matrix with the all-one row increases its rank by one. Therefore, the nullity of $\mathbf{A}(S)$ is $|S| - |V(S)|$. Hence, the dimension of $\mathcal{P}(S)$ is at most $|S| - |V(S)|$.

Next, we show that there exists a $\mathbf{u} > \mathbf{0}$ such that $\mathbf{A}(S)\mathbf{u} = \mathbf{b}$. Since the nullity of $\mathbf{B}(D(S))$ is positive, there exists a $\mathbf{u}'$ such that $\mathbf{A}(S)\mathbf{u}' = \mathbf{b}$. Since $D(S)$ is strongly connected, we apply Lemma 4.1 to find a vector $\mathbf{v} > \mathbf{0}$ such that $\mathbf{A}(S)\mathbf{v} = \mu\mathbf{b}$ for some positive μ. Choose μ' sufficiently large so that $\mathbf{u}' + \mu'\mathbf{v} > \mathbf{0}$ and set $\mathbf{u} = (\mathbf{u}' + \mu'\mathbf{v})/(1 + \mu'\mu)$. One can easily verify that $\mathbf{A}(S)\mathbf{u} = \mathbf{b}$.

To complete the proof, we exhibit a set of $|S| - |V(S)| + 1$ affinely independent points in $\mathcal{P}(S)$. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{|S|-|V(S)|}$ be linearly independent vectors that span the null space of $\mathbf{A}(S)$. Since $\mathbf{u}$ has strictly positive entries, we can find ϵ small enough so that $\mathbf{u} + \epsilon\mathbf{u}_i$ belongs to $\mathcal{P}(S)$ for all $1 \leq i \leq |S| - |V(S)|$. Therefore $\{\mathbf{u}, \mathbf{u} + \epsilon\mathbf{u}_1, \mathbf{u} + \epsilon\mathbf{u}_2, \ldots, \mathbf{u} + \epsilon\mathbf{u}_{|S|-|V(S)|}\}$ is the desired set of $|S| - |V(S)| + 1$ affinely independent points in $\mathcal{P}(S)$. □

Lemma 5.4. Suppose $D(S)$ is strongly connected. Then $\mathcal{P}^{\circ}(S) = \{\mathbf{u} \in \mathbb{R}^{|S|} : \mathbf{A}(S)\mathbf{u} = \mathbf{b}, \mathbf{u} > \mathbf{0}\}$. Therefore, $|\mathcal{LP}_{>0}(n; S)| = L_{\mathcal{P}^{\circ}(S)}(n - \ell + 1)$.

Proof: Let $\mathbf{u} > \mathbf{0}$ be such that $\mathbf{A}(S)\mathbf{u} = \mathbf{b}$. For any $\mathbf{u}' \in \mathcal{P}(S)$, we have $\mathbf{A}(S)\mathbf{u}' = \mathbf{b}$ and hence, $\mathbf{A}(S)(\mathbf{u} - \mathbf{u}') = \mathbf{0}$. Since $\mathbf{u}$ has strictly positive entries, we choose ϵ small enough so that $\mathbf{u} + \epsilon(\mathbf{u} - \mathbf{u}') \geq \mathbf{0}$. Therefore, $\mathbf{u} + \epsilon(\mathbf{u} - \mathbf{u}')$ belongs to $\mathcal{P}(S)$ and $\mathbf{u}$ belongs to the interior of $\mathcal{P}(S)$.

Conversely, let $\mathbf{u} \in \mathcal{P}(S)$, with $u_{\mathbf{z}} = 0$ for some $\mathbf{z} \in S$. Since $D(S)$ is strongly connected, from the proof of Lemma 5.3, there exists a $\mathbf{u}' \in \mathcal{P}(S)$ with $\mathbf{u}' > \mathbf{0}$. Hence, for all $\epsilon > 0$, the $\mathbf{z}$-coordinate of $\mathbf{u} + \epsilon(\mathbf{u} - \mathbf{u}')$ is given by $-\epsilon u'_{\mathbf{z}}$, which is always negative. In other words, $\mathbf{u}$ does not belong to $\mathcal{P}^\circ(S)$. □

Therefore, using Ehrhart's theorem and Ehrhart-Macdonald reciprocity along with Lemmas 5.3 and 5.4, we arrive at the fact that $|\mathcal{LP}_{>0}(n; S)|$ and $|\mathcal{LP}_{\geq 0}(n; S)|$ are quasipolynomials in n whose coefficients are periodic in n.

In order to determine the period of the quasipolynomials, we characterize the vertex set of $\mathcal{P}(S)$.

Definition 5.1. A point $\mathbf{v}$ in a polytope is a *vertex* if $\mathbf{v}$ cannot be expressed as a convex combination of the other points.

Lemma 5.5. The vertex set of $\mathcal{P}(S)$ is given by $\{\boldsymbol{\chi}(C)/|C| : C \text{ is a cycle in } D(S)\}$.

Proof: First, observe that $\boldsymbol{\chi}(C)/|C|$ belongs to $\mathcal{P}(S)$ for any cycle C in $D(S)$.

Let $\mathbf{v} \in \mathcal{P}(S)$ and suppose that $\mathbf{v}$ is a vertex. Since $\mathcal{P}(S)$ is rational, its vertex $\mathbf{v}$ has rational coordinates (see [24, Section 2.8, Appendix A]). Let μ be a positive integer such that $\mu\mathbf{v}$ has integer entries. Construct the multigraph D' on $V(S)$ by adding $\mu v_{\mathbf{z}}$ copies of the arc $\mathbf{z}$ for all $\mathbf{z} \in S$. Since $\mathbf{v} \in \mathcal{P}(S)$, we have $\mathbf{B}(S)\mu\mathbf{v} = \mathbf{0}$ and hence, each of the strongly connected components of D' are Eulerian. Therefore, the arc set of D' can be decomposed into disjoint cycles. In other words, for some cycles $C_1, C_2, \ldots, C_t$, we have

$$\mu\mathbf{v} = \sum_{t=1} \boldsymbol{\chi}(C_t), \quad \text{that is,} \quad \mathbf{v} = \sum_{t=1} \frac{|C_t|}{\mu} \frac{\boldsymbol{\chi}(C_t)}{|C_t|}.$$

Since $\mathbf{v}$ is a vertex, $\mathbf{v}$ cannot be expressed as a convex combination of the other points. So, $t = 1$ and hence, $\mathbf{v} = \boldsymbol{\chi}(C)/|C|$ for some cycle C.

Conversely, we show that for any cycle C in $D(S)$, $\boldsymbol{\chi}(C)/|C|$ cannot be expressed as a convex combination of other points in $\mathcal{P}(S)$. Suppose otherwise. Then there exist cycles $C_1, C_2, \ldots, C_t$ distinct from C and nonnegative scalars $\alpha_1, \alpha_2, \ldots, \alpha_t$ such that $\boldsymbol{\chi}(C) = \sum_{i=1}^{t} \alpha_i \boldsymbol{\chi}(C_i)$. For each j, let e_j be an arc that belongs to C_j but not C. Then

$$0 = \boldsymbol{\chi}(C)_{e_j} = \sum_{1 \leq i \leq t} \alpha_i \boldsymbol{\chi}(C_i)_{e_j} \geq \alpha_j \boldsymbol{\chi}(C_j)_{e_j} = \alpha_j.$$

Hence, we have that $\alpha_j = 0$ for all j. Therefore, $\boldsymbol{\chi}(C) = \mathbf{0}$, a contradiction. □

Let $\lambda_S = \operatorname{lcm}\{|C| : C \text{ is a cycle in } D(S)\}$, where lcm denotes the lowest common multiple. Then the period of the quasipolynomial $L_{\mathcal{P}(S)}(n - \ell + 1)$ divides λ_S by Ehrhart's theorem.

Let us dilate the polytope $\mathcal{P}(S)$ by λ_S and consider the polytope $\lambda_S\mathcal{P}(S)$ and $L_{\lambda_S\mathcal{P}(S)}(t)$. 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,

$$|\bar{\mathcal{Q}}(n; S)| \geq L_{\lambda_S\mathcal{P}^\circ(S)}(t) = \Omega\left(t^{|S|-|V(S)|}\right),$$

whenever $n - \ell + 1 = \lambda_S t$ or $\lambda_S | (n - \ell + 1)$, and therefore, $|\bar{\mathcal{Q}}(n; S)| = \Theta'\left(n^{|S|-|V(S)|}\right)$. This completes the proof of Theorem 4.4.

In the special case where $D(S)$ contains a loop, we can show further that the leading coefficients of the quasipolynomials $|\mathcal{LP}_{>0}(n; S)|$ and $|\mathcal{LP}_{\geq 0}(n; S)|$ are the same and constant. This result is a direct consequence of Ehrhart-Macdonald reciprocity and the fact that $|\mathcal{LP}_{>0}(n; [\![q]\!]^\ell)|$ is monotonically increasing. We demonstrate the latter claim in Appendix C.

Note that when $S = [\![q]\!]^\ell$, Corollary 5.6 yields (7), a result of Jacquet *et al.* [15].

Corollary 5.6. Suppose $D(S)$ is strongly connected. If $D(S)$ contains a loop, then

$$\begin{aligned} |\mathcal{LP}_{>0}(n; S)| \sim |\bar{\mathcal{Q}}(n; S)| &\sim |\mathcal{LP}_{\geq 0}(n; S)| \\ &\sim c(S)n^{|S|-|V(S)|} + O(n^{|S|-|V(S)|-1}), \end{aligned} \tag{9}$$

for some constant $c(S)$.

6. Constructive Lower Bounds

Fix $S \subseteq [\![q]\!]^\ell$ and recall that $\mathbf{p}\mathcal{Q}(n; S)$ denotes the set of all ℓ-gram profile vectors of words in $\mathcal{Q}(n; S)$. For ease of exposition, we henceforth identify words in $\mathcal{Q}(n; S)$ with their corresponding profile vectors in $\mathbf{p}\mathcal{Q}(n; S)$. In Section 8, we provide an efficient method to map a profile vector in $\mathbf{p}\mathcal{Q}(n; S)$ back to a q-ary codeword in $\mathcal{Q}(n; S)$, Therefore, in this section, we construct GRCs as sets of profile vectors $\mathbf{p}\mathcal{Q}(n; S)$ which we may map back to corresponding q-ary codewords in $\mathcal{Q}(n; S)$.

Suppose that $\mathcal{C}$ is an (N, d)-AECC. We construct GRCs from $\mathcal{C}$ via the following methods:

(i) When $N = |S|$, we intersect $\mathcal{C}$ with $\mathbf{p}\mathcal{Q}(n; S)$ to obtain an ℓ-gram reconstruction code. In other words, we pick out the codewords in $\mathcal{C}$ that are also profile vectors. Specifically, $\mathcal{C} \cap \mathbf{p}\mathcal{Q}(n; S)$ is an $(n, d; S)$-GRC. However, the size $|\mathcal{C} \cap \mathbf{p}\mathcal{Q}(n; S)|$ is usually smaller than $|\mathcal{C}|$ and so, we provide estimates to $|\mathcal{C} \cap \mathbf{p}\mathcal{Q}(n; S)|$ for a classical family of AECCs in Section 6-A.

(ii) When $N < |S|$, we extend each codeword in $\mathcal{C}$ to a profile vector of length $|S|$ in $\mathbf{p}\mathcal{Q}(n; S)$. In contrast to the previous construction, we may in principle obtain an $(n, d; S)$-GRC with the same cardinality as $\mathcal{C}$. However, one may not always be able to extend an arbitrary word to a profile vector. Section 6-B describes one method of mapping words in $[\![m]\!]^N$ to $\mathbf{p}\mathcal{Q}(n; q, \ell)$ that preserves the code size for a suitable choice of the parameters m and N. In addition, this mapping also preserves the distance of the original code $\mathcal{C}$.

A. Intersection with $\mathbf{p}\mathcal{Q}(n; S)$

In this section, we assume $N = |S|$ and we estimate $|\mathcal{C} \cap \mathbf{p}\mathcal{Q}(n; S)|$ when $\mathcal{C}$ belongs to a classical family of AECCs proposed by Varshamov [29]. 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 $\boldsymbol{\beta} \in (\mathbb{Z}/p\mathbb{Z})^d$ and define the code

$$\mathcal{C}(\mathbf{H}, \boldsymbol{\beta}) \triangleq \{\mathbf{u} : \mathbf{H}\mathbf{u} \equiv \boldsymbol{\beta} \bmod p\}. \tag{10}$$

Then, $\mathcal{C}(\mathbf{H}, \boldsymbol{\beta})$ is an $(N, d+1)$-AECC [29]. Hence, $\mathcal{C}(\mathbf{H}, \boldsymbol{\beta}) \cap \mathbf{p}\mathcal{Q}(n; S)$ is an $(n, d+1; S)$-GRC for all $\boldsymbol{\beta} \in (\mathbb{Z}/p\mathbb{Z})^d$. Therefore, by the pigeonhole principle, there exists a $\boldsymbol{\beta}$ such that $|\mathcal{C}(\mathbf{H}, \boldsymbol{\beta}) \cap \mathbf{p}\mathcal{Q}(n; S)|$ is at least $|\mathbf{p}\mathcal{Q}(n; S)|/p^d$. However, the choice of $\boldsymbol{\beta}$ that guarantees this lower bound is not known.

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 $\mathcal{C}(\mathbf{H}, \boldsymbol{\beta}) \cap \mathbf{p}\mathcal{Q}(n; S)$ as a function of n. As before, instead of looking at $\mathbf{p}\mathcal{Q}(n; S)$ directly, we consider the set of closed words $\bar{\mathcal{Q}}(n; S)$ and the corresponding set of profile vectors $\mathbf{p}\bar{\mathcal{Q}}(n; S)$.

Let $\boldsymbol{\beta} = \mathbf{0}$ and choose $\mathbf{H}$ and p based on the restricted de Bruijn digraph $D(S)$. For an arbitrary matrix $\mathbf{M}$, let $\mathrm{Null}_{>\mathbf{0}}\mathbf{M}$ denote the set of vectors in the null space of $\mathbf{M}$ that have positive entries. We assume $D(S)$ to be strongly connected so that $\mathrm{Null}_{>\mathbf{0}}\mathbf{B}(D(S))$ is nonempty from Lemma 4.1. Hence, we choose $\mathbf{H}$ and p such that $\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathrm{Null}_{>\mathbf{0}}\mathbf{B}(D(S))$ is nonempty.

Define the $(|V(S)| + 1 + d) \times (|S| + d)$-matrix

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

where $\mathbf{A}(S)$ is as described in Section 4. Let $\mathbf{b}$ be a vector of length $|V(S)| + 1 + d$ that has 1 as the first entry and zeros elsewhere, and define the polytope

$$\mathcal{P}_{\mathrm{GRC}}(\mathbf{H}, S) \triangleq \{\mathbf{u} \in \mathbb{R}^{|S|+d} : \mathbf{A}(\mathbf{H}, S)\mathbf{u} = \mathbf{b}, \mathbf{u} \geq \mathbf{0}\} \tag{11}$$

Since $\mathcal{LP}_{>0}(n; S) \subseteq \mathbf{p}\bar{\mathcal{Q}}(n; S) \subseteq \mathbf{p}\mathcal{Q}(n; S)$, $|\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathcal{LP}_{>0}(n; S)|$ is a lower bound for $|\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathbf{p}\mathcal{Q}(n; S)|$. The following proposition demonstrates that $|\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathcal{LP}_{>0}(n; S)|$ is given by the number of lattice points in the interior of a dilation of $\mathcal{P}_{\mathrm{GRC}}(\mathbf{H}, S)$.

Proposition 6.1. Let $\mathcal{C}(\mathbf{H}, \mathbf{0})$ and $\mathcal{P}_{\mathrm{GRC}}(\mathbf{H}, S)$ be defined as above. If $D(S)$ is strongly connected and $\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathrm{Null}_{>\mathbf{0}}\mathbf{B}(D(S))$ is nonempty, then $|\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathcal{LP}_{>0}(n; S)| = \left|\mathbb{Z}^{N+d} \cap (n - \ell + 1)\mathcal{P}^{\circ}_{\mathrm{GRC}}(\mathbf{H}, S)\right|$.

Proof: Similar to Lemma 5.4, we have that $\mathcal{P}^{\circ}_{\mathrm{GRC}}(\mathbf{H}, S) = \{\mathbf{u} \in \mathbb{R}^{|S|+d} : \mathbf{A}(\mathbf{H}, S)\mathbf{u} = \mathbf{b}, \mathbf{u} > \mathbf{0}\}$, and we defer the proof of this claim to Appendix D.

To prove the desired sets have the same cardinality, we construct a bijection between the two maps. Let $\mathbf{u} > \mathbf{0}$ be such that $\mathbf{A}(\mathbf{H}, S)\mathbf{u} = (n - \ell + 1)\mathbf{b}$. Let $\mathbf{u} = (\mathbf{u}_0, \boldsymbol{\beta}')$, where the vector $\mathbf{u}_0$ is the vector $\mathbf{u}$ restricted to the first N coordinates and $\boldsymbol{\beta}'$ is the vector $\mathbf{u}$ restricted to the last d coordinates. Then $\mathbf{A}(S)\mathbf{u}_0 = (n - \ell + 1)\mathbf{b}_0$, where $\mathbf{b}_0$ is a vector of length $|V(S)| + 1$ with one in its first coordinate and zeros elsewhere. Hence, $\mathbf{u}_0 \in \mathcal{LP}_{>0}(n; S)$. On the other hand, $\mathbf{H}\mathbf{u}_0 = p\boldsymbol{\beta}'$ and so, $\mathbf{H}\mathbf{u}_0 \equiv \mathbf{0} \bmod p$, implying that $\mathbf{u}_0 \in \mathcal{C}(\mathbf{H}, \mathbf{0})$. Therefore, $\phi(\mathbf{u}) = \mathbf{u}_0$ is well-defined map from $\{\mathbf{u} \in \mathbb{Z}^{N+d} : \mathbf{A}(\mathbf{H}, S)\mathbf{u} = (n - \ell + 1)\mathbf{b} \text{ and } \mathbf{u} > \mathbf{0}\}$ to $\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathcal{LP}_{>0}(n; S)$.

Next, consider $\mathbf{u}_0 \in \mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathcal{LP}_{>0}(n; S)$. Then $\mathbf{A}(S)\mathbf{u}_0 = (n - \ell + 1)\mathbf{b}_0$. Also, $\mathbf{H}\mathbf{u}_0 \equiv \mathbf{0} \bmod p$ and hence, $\frac{1}{p}\mathbf{H}\mathbf{u}_0$ has integer coordinates. Then $\psi(\mathbf{u}_0) = (\mathbf{u}_0, \frac{1}{p}\mathbf{H}\mathbf{u}_0)$ is a well-defined map from $\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathcal{LP}_{>0}(n; S)$ to $\{\mathbf{u} \in \mathbb{Z}^{N+d} : \mathbf{A}(\mathbf{H}, S)\mathbf{u} = (n - \ell + 1)\mathbf{b} \text{ and } \mathbf{u} > \mathbf{0}\}$.

Finally, to demonstrate that both ϕ and ψ are bijections, we verify that $\psi \circ \phi$ and $\phi \circ \psi$ are both identity maps on $\{\mathbf{u} \in \mathbb{Z}^{N+d} : \mathbf{A}(\mathbf{H}, S)\mathbf{u} = (n - \ell + 1)\mathbf{b} \text{ and } \mathbf{u} > \mathbf{0}\}$ and $\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathcal{LP}_{>0}(n; S)$, respectively. Indeed,

$$\psi \circ \phi((\mathbf{u}_0, \boldsymbol{\beta}')) = \psi(\mathbf{u}_0) = (\mathbf{u}_0, \frac{1}{p}\mathbf{H}\mathbf{u}_0) = (\mathbf{u}_0, \boldsymbol{\beta}'),$$

$$\phi \circ \psi(\mathbf{u}_0) = \psi((\mathbf{u}_0, \frac{1}{p}\mathbf{H}\mathbf{u}_0)) = \mathbf{u}_0.$$

Hence, the two sets have the same cardinality. □

As before, we compute the dimension of $\mathcal{P}_{\mathrm{GRC}}(\mathbf{H}, S)$ and characterize its vertex set. Since the proofs are similar to the ones in Section 5, the reader is referred to Appendix D for a detailed analysis.

Lemma 6.2. Let $\mathcal{C}(\mathbf{H}, \mathbf{0})$ and $\mathcal{P}_{\mathrm{GRC}}(\mathbf{H}, S)$ be defined as above. Suppose further that $D(S)$ is strongly connected and $\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathrm{Null}_{>\mathbf{0}}\mathbf{B}(D(S))$ is nonempty. The dimension of $\mathcal{P}_{\mathrm{GRC}}(\mathbf{H}, S)$ is $|S| - |V(S)|$, while its vertex set is given by

$$\left\{ \left(\frac{\boldsymbol{\chi}(C)}{|C|}, \frac{\mathbf{H}\boldsymbol{\chi}(C)}{p|C|} \right) : C \text{ is a cycle in } D(S) \right\}.$$

Let $\lambda_{\mathrm{GRC}} = \mathrm{lcm}\{|C| : C \text{ is a cycle in } D(S)\} \cup \{p\}$. Then Lemma 6.2, Ehrhart's theorem and Ehrhart-Macdonald's reciprocity imply that $L_{\mathcal{P}^{\circ}_{\mathrm{GRC}}(\mathbf{H},S)}(t)$ is a quasipolynomial of degree $|S| - |V(S)|$ whose period divides λ_{GRC}. As in Section 5, we dilate the polytope $\mathcal{P}_{\mathrm{GRC}}(\mathbf{H}, S)$ by λ_{GRC} to obtain an integer polytope and assume that the polynomial $L_{\lambda_{\mathrm{GRC}}\mathcal{P}_{\mathrm{GRC}}(\mathbf{H},S)}(t)$ has leading coefficient c. Hence, whenever $n - \ell + 1 = \lambda_{\mathrm{GRC}}t$, that is, whenever $\lambda_{\mathrm{GRC}} | (n - \ell + 1)$,

$$\begin{aligned} &|\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathcal{LP}_{>0}(n; S)| \\ &= L_{\lambda_{\mathrm{GRC}}\mathcal{P}^{\circ}_{\mathrm{GRC}}(\mathbf{H},S)}(t) \\ &= ct^{|S|-|V(S)|} + O(t^{|S|-|V(S)|-1}) \\ &= c(n/\lambda_{\mathrm{GRC}})^{|S|-|V(S)|} + O(n^{|S|-|V(S)|-1}). \end{aligned}$$

We denote $c/\lambda_{\mathrm{GRC}}^{|S|-|V(S)|}$ by $c(\mathbf{H}, S)$ and summarize the results in the following theorem.

Theorem 6.3. Fix $S \subseteq [\![q]\!]^{\ell}$ and d. Choose $\mathbf{H}$ and p so that $\mathcal{C}(\mathbf{H}, \mathbf{0})$ is an $(|S|, d + 1)$-AECC and $\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathrm{Null}_{>\mathbf{0}}\mathbf{B}(D(S))$ is nonempty. Suppose that $\lambda_{\mathrm{GRC}} = \mathrm{lcm}\{\{|C| : C \text{ is a cycle in } D(S)\} \cup \{p\}\}$. Then there exists a constant $c(\mathbf{H}, S)$ such that whenever $\lambda_{\mathrm{GRC}} | (n - \ell + 1)$,

$$|\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathbf{p}\mathcal{Q}(n; S)| \geq c(\mathbf{H}, S)n^{|S|-|V(S)|} + O(n^{|S|-|V(S)|-1}).$$

Hence, it follows from Theorem 6.3, we have $C(n, d; S) = \Omega'(n^{|S|-|V(S)|})$ when d is constant. Since $C(n, d; S) \leq |\mathcal{Q}(n; S)| = O(n^{|S|-|V(S)|})$, we have $C(n, d; S) = \Theta'(n^{|S|-|V(S)|})$.

In Example 7.1, we consider the case $S = [\![2]\!]^3$ and $d = 2$. We then computed the number of codewords in $\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \mathcal{LP}_{>0}(n; 2, 3)$ to be $12168t^4 - 1248t^3 + 131t^2 - 16t + 1$, where

$n = 156t + 2$. In other words, we have that $C(n, 3; [\![2]\!]^3) = \Theta'(n^4)$. Details are provided in Section 7.

Remark 2.

(i) We consider the complexity of determining p and H. The value of p may be determined in time polynomial in N since there always exists a prime number between N and $2N$ by Bertrand's postulate [30] and the running time of a primality test is polynomial in $\log N$ [31]. The construction $\mathbf{H}$ can be completed in time polynomial in N, since multiplication in the field $\mathbb{F}_p$ has time complexity polynomial in $\log N$ and there are dN entries to fill in $\mathbf{H}$.

(ii) When $D(S)$ is Eulerian, the all-ones vector belongs to $\mathrm{Null}_{>\mathbf{0}}\mathbf{B}(D(S))$. It then suffices to construct a check matrix $\mathbf{H}$ such that the elements in each row sum to zero. To this end, one may simply choose p such that N divides $p-1$ and then pick the N field elements $\alpha_1, \alpha_2, \ldots, \alpha_N$ so that $\alpha_j^N = 1$ for all $1 \le j \le N$. It is easy to check that one has $\sum_{j=1}^N \alpha_j^i = 0 \bmod p$ for all $1 \le i \le d < N$.
So, one can construct $\mathbf{H}$ efficiently whenever $D(S)$ is Eulerian. As observed from Proposition 4.2, we have that $D(q, \ell; q^*, [w_1, w_2])$, the graphs that we work with, are Eulerian.

(iii) For clarity of exposition, we considered the case where $\boldsymbol{\beta} = \mathbf{0}$. For general $\boldsymbol{\beta}$, similar results hold with suitable modifications to $\mathbf{b}$ and the polytope $\mathcal{P}_{\mathrm{GRC}}(\mathbf{H}, S)$ defined in (11). In particular, we can set the last d entries of $\mathbf{b}$ to be $\boldsymbol{\beta}$ and then Proposition 6.1 can be modified to show that $|\mathcal{C}(\mathbf{H}, (n-\ell+1)\boldsymbol{\beta}) \cap \mathcal{LP}_{>0}(n; S)|$ is given by the number of lattice points in the interior of a dilation of the new polytope $\mathcal{P}_{\mathrm{GRC}}(\mathbf{H}, S)$. Observe that $\mathcal{C}(\mathbf{H}, (n-\ell+1)\boldsymbol{\beta}) \cap \mathcal{LP}_{>0}(n; S)$ remains an $(|S|, d+1)$-AECC and corresponding versions of Lemma 6.2 and Theorem 6.3 follow using similar derivations.

B. Systematic Encoding of Profile Vectors

In this subsection, we look at efficient one-to-one mappings from $[\![m]\!]^N$ to $\mathbf{p}\mathcal{Q}(n; S)$. As with usual constrained coding problems, we are interested in maximizing the number of messages, i.e. the size of m^N, so that the number of messages is close to $|\mathbf{p}\mathcal{Q}(n; S)| = \Theta'(n^{|S|-|V(S)|})$. We achieve this goal by exhibiting a systematic encoder with $m = \Theta(n)$ and $N = |S| - |V(S)| - 1$. More formally, we prove the following theorem.

Theorem 6.4 (Systematic Encoder)**.** Fix n and $S \subseteq [\![q]\!]^\ell$. Pick any m so that

$$m \le \frac{n - \ell + 1}{\binom{|V(S)|}{2}(q-1) + |S| - |V(S)| - 1}. \tag{12}$$

Suppose further that $D(S)$ is Hamiltonian and contains a loop. Then, there exists a set $I \subseteq S$ of coordinates of size $|S| - |V(S)| - 1$ with the following property: for any $\mathbf{v} \in [\![m]\!]^I$, there exists an ℓ-gram profile vector $\mathbf{u} \in \mathbf{p}\mathcal{Q}(n; S)$ such that $\mathbf{u}|_I = \mathbf{v}$. Furthermore, $\mathbf{u}$ can be found in time $O(|V(S)|)$.

In other words, given any word $\mathbf{v}$ of length $N = |I| = |S| - |V(S)| - 1$, one can always extend it to obtain a profile vector $\mathbf{u} \in \mathbf{p}\mathcal{Q}(n; S)$ of length $|S|$. As pointed out earlier, this theorem provides a simple way of constructing ℓ-gram codes from AECCs and we sketch the construction in what follows.

Let $\phi_{\mathrm{sys}}(\mathbf{v})$ denote the profile vector resulting from Theorem 6.4 given input $\mathbf{v}$. Consider an m-ary (N, d)-AECC $\mathcal{C}$ with $N = |S| - |V(S)| - 1$ and m satisfying (12). Let $\phi_{\mathrm{sys}}(\mathcal{C}) \triangleq \{\phi_{\mathrm{sys}}(\mathbf{v}) : \mathbf{v} \in \mathcal{C}\}$. Then $\phi_{\mathrm{sys}}(\mathcal{C}) \subseteq \mathbf{p}\mathcal{Q}(n; S)$. Furthermore, $\phi_{\mathrm{sys}}(\mathcal{C})$ has asymmetric distance at least d since restricting the code $\phi_{\mathrm{sys}}(\mathcal{C})$ on the coordinates in I yields $\mathcal{C}$. Hence, we have the following corollary.

Corollary 6.5. Fix n and $S \subseteq [\![q]\!]^\ell$ and pick m satisfying (12). Suppose $D(S)$ is Hamiltonian and contains a loop. If $\mathcal{C}$ is an m-ary $(|S| - |V(S)| - 1, d)$-AECC, then $\phi_{\mathrm{sys}}(\mathcal{C}) \triangleq \{\phi_{\mathrm{sys}}(\mathbf{v}) : \mathbf{v} \in \mathcal{C}\}$ is a $(n, d; S)$-GRC.

For compactness, we write V, $\mathbf{A}$ and $\mathbf{B}$, instead of $V(S)$, $\mathbf{A}(S)$ and $\mathbf{B}(D(S))$. To prove Theorem 6.4, consider the restricted de Bruijn digraph $D(S)$. By the assumptions of the theorem, denote the set of $|V|$ arcs in a Hamiltonian cycle as H and the arc corresponding to a loop by $\mathbf{a}_0$. We set I to be $S \setminus (H \cup \{\mathbf{a}_0\})$.

We reorder the coordinates so that the arcs in H are ordered first, followed by the arc $\mathbf{a}_0$ and then the arcs in I. So, given $\mathbf{v} = (v_1, v_2, \ldots, v_{|I|}) \in [\![m]\!]^{|I|}$, the proof of Theorem 6.4 essentially reduces to finding integers $x_1, x_2, \ldots, x_{|V|}, y$ such that

$$\mathbf{A}\left(x_1, x_2, \ldots, x_{|V|}, y, v_1, v_2, \ldots, v_{|I|}\right)^T = (n - \ell + 1)\mathbf{b}. \tag{13}$$

Considering the first row of $\mathbf{A}$ separately from the remaining rows, we see that (13) is equivalent to the following system of equations:

$$\sum_{i=1}^{|V|} x_i + y = (n - \ell + 1) - \sum_{i=1}^{|I|} v_i, \tag{14}$$

$$\mathbf{0} = \mathbf{B}\begin{pmatrix} x_1 \\ \vdots \\ x_{|V|} \\ y \\ v_1 \\ \vdots \\ v_{|I|} \end{pmatrix} = \mathbf{B}\begin{pmatrix} x_1 \\ \vdots \\ x_{|V|} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \mathbf{B}\begin{pmatrix} 0 \\ \vdots \\ 0 \\ y \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \mathbf{B}\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ v_1 \\ \vdots \\ v_{|I|} \end{pmatrix}. \tag{15}$$

Since the first $|V|$ columns of $\mathbf{B}$ correspond to the arcs in H, we have

$$\mathbf{B}\left(x_1, \ldots, x_{|V|}, 0, 0, \ldots, 0\right)^T = \begin{pmatrix} x_2 - x_1 \\ x_3 - x_2 \\ \vdots \\ x_1 - x_{|V|} \end{pmatrix}.$$

Since the $(|V| + 1)$-th column of $\mathbf{B}$ is a $\mathbf{0}$-column, we have $\mathbf{B}(0, \ldots, 0, y, 0, \ldots, 0)^T = \mathbf{0}$ for any y.

For the final summand, let $\mathbf{B}\left(0, \ldots, 0, 0, v_1, \ldots, v_{|I|}\right)^T = (r_1, r_2, \ldots, r_{|V|})^T$. We can then rewrite (15) as

$$x_i - x_{i+1} = r_i, \text{ for } 1 \le i \le |V| - 1, \text{ and } x_{|V|} - x_1 = r_{|V|}. \tag{16}$$

Since $\mathbf{1}^T\mathbf{B} = \mathbf{0}^T$, we have $\mathbf{1}^T(r_1, r_2, \ldots, r_{|V|})^T = \sum_{i=1}^{|V|} r_i = 0$. Furthermore, we assume without loss of generality that $\sum_{i=1}^{j} r_i \geq 0$, for all $1 \leq j \leq |V|$. This can be achieved by cyclically relabelling the nodes and we prove this in Appendix E.

It suffices to show that an integer solution for (16) and (14) exists, satisfying $y \geq 1$ and $x_i \geq 1$ for $i \in [|V|]$. Consider the following choices of x_i and y:

$$x_i = 1 + \sum_{j=1}^{i-1} r_j,$$

$$y = (n - \ell + 1) - \sum_{i=1}^{|I|} v_i - \sum_{i=1}^{|V|} x_i.$$

Clearly, x_i and y satisfy (14) and (16). Since each v_i is an integer, all r_i are integers, so x_i and y are also integers. Furthermore, each $x_i \geq 1$, since we chose the labeling so that $\sum_{j=1}^{i-1} r_j \geq 0$. We still must show that $y \geq 1$.

First, we observe that $r_i < (q-1)m$ for all i, since each node has at most $(q-1)$ incoming arcs in I and by design, each v_i is strictly less than m. Thus, each x_i satisfies

$$x_i < 1 + (i-1)(q-1)m.$$

Summing over all i, we have

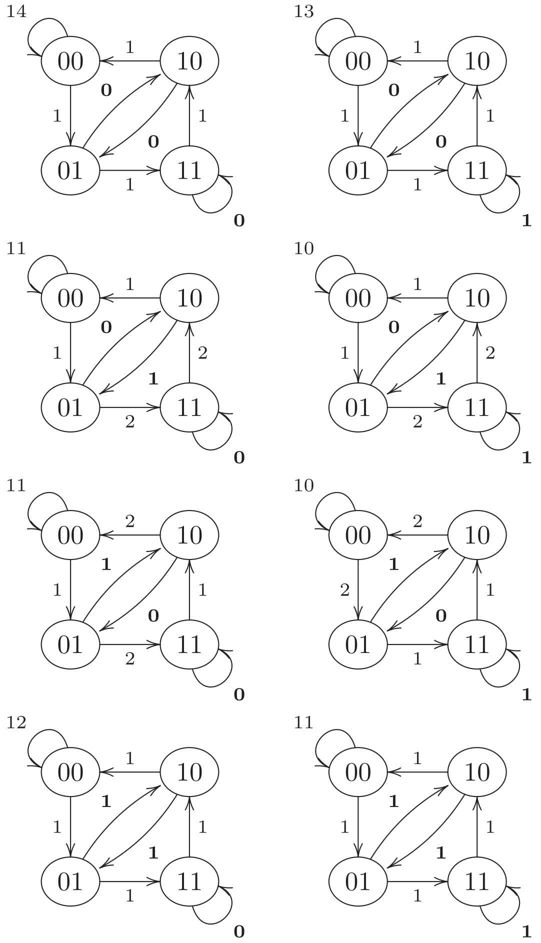
$$\sum_{i=1}^{|V|} x_i \leq \sum_{i=1}^{|I|} (i-1)(q-1)m = (q-1)m\binom{|V|}{2}.$$

Since also each $v_i \leq m$, we have

$$y \geq (n - \ell + 1) - m\left[|I| + (q-1)\binom{|V|}{2}\right].$$

By the choice of m, it follows that $y \geq 0$. This completes the proof of Theorem 6.4.

Example 6.1. Let $S = [\![2]\!]^3$ and let $n = 20$. Then Theorem 6.4 states that there is a systematic encoder that maps words from $[\![2]\!]^3$ into $\mathbf{p}\mathcal{Q}(20; 2, 3)$. Following the convention in Fig. 2 and Example 4.2, we list all eight encoded profile vectors (as edge labellings on $D([\![2]\!]^3)$) with their **systematic components** highlighted in boldface.



For instance, the codeword $000 \in [\![2]\!]^3$ is mapped to the profile vector $(14, 1, \mathbf{0}, 1, 1, \mathbf{0}, 1, \mathbf{0})$. Via the EULER map described in Section 8, this profile vector is mapped to $00\cdots01100 \in \mathcal{Q}(20, 2, 3)$.

Observe that we can systematically encode $[\![2]\!]^3$ into $\mathbf{p}\mathcal{Q}(n; 2, 3)$ even when n is smaller than 20. In fact, in this example, we can systematically encode $[\![2]\!]^3$ into $\mathbf{p}\mathcal{Q}(10; 2, 3)$. In general, we can can systematically encode $[\![m]\!]^3$ into $\mathbf{p}\mathcal{Q}(4m + 2; 2, 3)$. In this case, the size of the message set is approximately $n^3/64$ while the number of all possible closed profile vectors is approximately $n^4/288$ [15].

In Section 7 and Example 7.1, we observe that the construction given in Section 6-A yields a larger code size. Nevertheless, the systematic encoder is conceptually simple and furthermore, the systematic property of the construction in Section 6-B can be exploited to integrate rank modulation codes into our coding schemes for DNA storage, useful for automatic decoding via *hybridization*. We describe this procedure in detail in Section 8.

7. NUMERICAL COMPUTATIONS FOR $S = S(q, \ell; q^*, [w_1, w_2])$

In what follows, we summarize numerical results for code sizes pertaining to the special case when $S = S(q, \ell; q^*, [w_1, w_2])$.

By Proposition 4.2, $D(q, \ell; q^*, [w_1, w_2])$ is Eulerian and therefore strongly connected. In other words, Theorem 4.4 applies and we have $|\mathcal{Q}(n; S)| = \Theta'(n^{|S|-|V(S)|})$, where $|S|$

is given by $|S(q,\ell;q^*,[w_1,w_2])| = \sum_{w=w_1}^{w_2}\binom{\ell}{w}(q^*)^w(q-q^*)^{\ell-w}$, while $|V(S)|$ is given by $|S(q,\ell-1;q^*,[w_1-1,w_2])| = \sum_{w=w_1-1}^{w_2}\binom{\ell-1}{w}(q^*)^w(q-q^*)^{\ell-1-w}$.

Let $D = |S| - |V(S)|$. We determine next the coefficient of n^D in $|\mathcal{Q}(n;S)|$. When $w_2 = \ell$, the digraph $D(q,\ell;q^*,[w_1,\ell])$ contains the loop that corresponds to the ℓ-gram $\mathbf{1}^T$. Hence, by Corollary 5.6, the desired coefficient is constant and we denote it by $c(q,\ell;q^*,[w_1,\ell])$. When $S = [\![q]\!]^\ell$, we denote this coefficient by $c(q,\ell)$ and remark that this value corresponds to the constant defined in Theorem 4.6.

When $w_2 < \ell$, the digraph $D(q,\ell;q^*,[w_1,w_2])$ does not contain any loops. Recall from Section 5 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 values for which λ_S divides $(n-\ell+1)$ or $n-\ell+1 = \lambda_S t$ for some integer t. In this instance, the desired coefficient is given by $c(q,\ell;q^*,[w_1,w_2]) \triangleq c/\lambda_S^D$, where c is the leading coefficient of the polynomial $L_{\lambda_S\mathcal{P}(S)}(t)$.

In summary, we have the following corollary.

Corollary 7.1. Consider $S = S(q,\ell;q^*,[w_1,w_2])$ and define

$$D = \sum_{w=w_1}^{w_2}\binom{\ell}{w}(q^*)^w(q-q^*)^{\ell-w} - \sum_{w=w_1-1}^{w_2}\binom{\ell-1}{w}(q^*)^w(q-q^*)^{\ell-1-w}.$$

Suppose that $\lambda_S = \mathrm{lcm}\{|C| : C \text{ is a cycle in } D(S)\}$. Then for some constant $c(q,\ell;q^*,[w_1,w_2])$,

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

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

When $S = [\![q]\!]^\ell$, we write $c(q,\ell)$ instead of $c(q,\ell;1,[0,\ell])$.

We determine $c(q,\ell;q^*,[w_1,w_2])$ 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.* [32], which is based on an algorithm of Barvinok [33]. Barvinok's algorithm essentially triangulates the supporting cones of the vertices of a polytope to obtain simplicial cones and then decompose the simplicial cones recursively into unimodular cones. As the rational generating functions of the resulting unimodular cones can be written down easily, adding and subtracting them according to the inclusion-exclusion principle and Brion's theorem gives the desired rational generating function of the polytope. The algorithm is shown to enumerate the number of lattice points in polynomial time when the dimension of the polytope is fixed.

Using `LattE`, we computed the desired coefficients for various values of $(q,\ell;q^*,[w_1,w_2])$. 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.* [15]), numerical computations of $c(q,\ell)$ and $c(q,\ell;q^*,[w_1,w_2])$ are feasible for certain moderate values of parameters. We tabulate these values in Table I and II.

TABLE I

COMPUTATION OF $c(q,\ell)$

q	ℓ	D	$c(q,\ell)$
2	2	2	1/4*
3	2	6	1/8640*
4	2	12	1/45984153600*
5	2	20	37/840810934025846784000000*
2	3	4	1/288*
3	3	18	887/358450977137334681600000
2	4	8	283/9754214400
2	5	16	722299813/94556837526637331349504000000

Entries marked by an asterisk refer to values that were also derived by Jacquet *et al.* [15].

Next, we provide numerical results for lower bounds on the code sizes derived in Section 6-A.

When $S = S(q,\ell;q^*,[w_1,w_2])$, the digraph $D(S)$ is Eulerian by Proposition 4.2 and hence, $\mathbf{1}$ belongs to $\mathrm{Null}_{>0}\mathbf{B}(D(S))$. Therefore, if $\mathcal{C}(\mathbf{H},\mathbf{0})$ contains the vector $\mathbf{1}$ as well, $\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathrm{Null}_{>0}\mathbf{B}(D(S))$ is nonempty and the condition of Theorem 6.3 is satisfied. Hence, we have the following corollary.

Corollary 7.2. Let $S = S(q,\ell;q^*,[w_1,w_2])$. Fix d and choose $\mathbf{H}$ and p such that $\mathcal{C}(\mathbf{H},\mathbf{0})$ is an $(|S|,d+1)$-AECC containing $\mathbf{1}$. Suppose that $\lambda_{\mathrm{GRC}} = \mathrm{lcm}\{\{|C| : C \text{ is a cycle in } D(S)\}\cup\{p\}\}$. Then there exists a constant $c(\mathbf{H},S)$ such that whenever $\lambda_{\mathrm{GRC}}|(n-\ell+1)$,

$$|\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathbf{p}\mathcal{Q}(n;S)| \geq c(\mathbf{H},S)n^D + O(n^{D-1}),$$

where $D = |S| - |V(S)| = \sum_{w=w_1}^{w_2}\binom{\ell}{w}(q^*)^w(q-q^*)^{\ell-w} - \sum_{w=w_1-1}^{w_2}\binom{\ell-1}{w}(q^*)^w(q-q^*)^{\ell-1-w}$.

Example 7.1. Let $S = [\![2]\!]^3$ and $d = 2$. 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 containing $\mathbf{1}$. We have $\lambda_{\mathrm{GRC}} = \mathrm{lcm}\{\{1,2,\ldots,8\}\cup\{13\}\} = 156$. Using `LattE`, we compute the lattice point enumerator of $\lambda_{\mathrm{GRC}}\mathcal{P}^{\circ}_{\mathrm{GRC}}(\mathbf{H},S)$ to be $12168t^4 - 1248t^3 + 131t^2 - 16t + 1$. Hence, for $n = 156t+2$, the number of codewords in $\mathcal{C}(\mathbf{H},\mathbf{0})\cap\mathcal{LP}_{>0}(n;2,3)$ is given by $12168t^4 - 1248t^3 + 131t^2 - 16t + 1$. When $t = 1$ or $n = 158$, there exist a $\left(158, 3; [\![2]\!]^3\right)$-GRC of size at least 11036.

We compare this result with the one provided by the construction using the systematic encoder described in Section 6-B and in particular, Example 6.1. When $n = 158$, we can systematically encode words in $[\![39]\!]^3$ into $\mathbf{p}\mathcal{Q}\left(158; [\![2]\!]^3\right)$. Hence, we consider a 39-ary $(3,3)$-AECC. Using Varshamov's construction with $p_1 = 5$ and $\mathbf{H}_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 4 & 4 \end{pmatrix}$, we obtain a 39-ary $(3,3)$-AECC of size 2368. Applying the systematic encoder in Theorem 6.4, we construct a $(158,3;2,3)$-GRC of size 2368.

TABLE II

COMPUTATION OF $c(q, \ell; q^*, [w_1, w_2])$. WE FIXED $q = 2$ AND $q^* = 1$.

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	3	5	15	5354228880	1106713336565579/739506679855711968646397952000000000
6	4	5	5	840	1/518400

TABLE III

COMPUTATIONS OF $c(\mathbf{H}, S)$

When $S = [\![2]\!]^3$, we have $c(2, 3) = 1/288$.

d	p	D	λ_{GRC}	$c(\mathbf{H}, S)$	$c(2, 3)/p^d$
1	11	4	132	1/3168	1/3168
2	13	4	156	1/48672	1/48672
3	13	4	156	1/632736	1/632736
4	17	4	204	1/24054048	1/24054048
5	17	4	204	1/24054048	1/408918816
6	17	4	204	1/24054048	1/6951619872

When $S = [\![2]\!]^4$, we have $c(2, 4) = 283/9754214400$.

d	p	D	λ_{GRC}	$c(\mathbf{H}, S)$	$c(2, 4)/p^d$
1	17	8	14280	283/165821644800	283/165821644800
2	17	8	14280	283/2818967961600	283/2818967961600
3	17	8	14280	283/47922455347200	283/47922455347200

When $S = S(2, 5; 1, [2, 3])$, we have $c(2, 5; 1, [2, 3]) = 1/5184000$.

d	p	D	λ_{GRC}	$c(\mathbf{H}, S)$	$c(2, 5; 1, [2, 3])/p^d$
1	23	6	2760	1/119232000	1/119232000
2	29	6	3480	1/4359744000	1/4359744000
3	29	6	3480	1/126432576000	1/126432576000

Using `LattE`, we determined $c(\mathbf{H}, S)$ for moderate parameter values and summarize the results in Table III.

We conclude this section with a conjecture on the relation between $c(q, \ell)$ and $c(\mathbf{H}, S)$.

Conjecture 7.3. *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 $c(q, \ell)$ and $c(\mathbf{H}, S)$ be the constants defined in Corollaries 7.1 and 7.2, respectively. Then $c(\mathbf{H}, S) \geq c(q, \ell)/p^d$.*

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

8. DECODING OF PROFILE VECTORS

Recall the DNA storage channel illustrated in Fig. 1. The channel takes as its input a word $\mathbf{x} \in \mathcal{Q}(n; S)$ and outputs a profile vector $\widehat{\mathbf{p}}(\mathbf{x}) \in \mathbb{Z}^{|S|}$. Assuming no errors, the vector $\widehat{\mathbf{p}}(\mathbf{x})$ corresponds to the correct profile vector $\mathbf{p}(\mathbf{x}; S) \in \mathbf{p}\mathcal{Q}(n; S)$. In this channel model and the code constructions in Section 6, we have implicitly assumed the existence of an efficient algorithm that decodes $\widehat{\mathbf{p}}(\mathbf{x}) \in \mathbb{Z}^{|S|}$ back to the message $\mathbf{x}$. We now describe this two-step algorithm in more detail.

The first step of decoding is to correct errors in $\widehat{\mathbf{p}}(\mathbf{x}) \in \mathbb{Z}^{|S|}$ to arrive at a profile vector of the valid codeword $\mathbf{p}(\mathbf{x}; S) \in \mathbf{p}\mathcal{Q}(n; S)$. For this purpose, one can use the conceptually simple Varshamov's decoding algorithm described in [18]. The algorithm reduces to recursive computations of residues of the channel output profile vectors with respect to the rows of the matrix $\mathbf{H}$ defining the code in (10) and solving a system of equations over a finite field.

The second step of decoding consists of converting the corrected profile vector into the corresponding codeword. For the purpose of describing this process, let $\mathbf{u}$ be a profile vector in $\mathbf{p}\mathcal{Q}(n; S)$ so that $\mathbf{u} = \mathbf{p}(\mathbf{x}; S)$ for some $\mathbf{x} \in \mathcal{Q}(n; S)$. As it was done in the proof of Lemma 4.3, we construct a multigraph on the node set $V(S)$ by adding $u_{\mathbf{z}}$ arcs for each $\mathbf{z} \in V(S)$. We remove any isolated nodes to arrive at a connected Eulerian multidigraph. We subsequently apply any linear-time algorithm like Hierholzer's algorithm [34] to this multidigraph to obtain an Eulerian walk. Hierholzer's algorithm uses two straightforward search steps:

- One starts by choosing a starting node in the multidigraph v and then proceeds by following a connected sequence of edges until returning to v. Note that the multidigraph is Eulerian so such a closed path will always exist. Note that one closed path may not cover all edges (or nodes) in the graph.
- If the path does not cover all edges, as long as there exists a node u on the last identified closed path that has emanating edges terminating in nodes not on the closed path, initiate another closed walk from the node u that does not share any edges with the current closed path. Merge the current path with the path initiated from u.

Most implementations of the Hierholzer's algorithm involve an arbitrary choice for the starting node and the subsequent nodes to visit. Hence, it is possible for the algorithm to produce different walks based on the same multigraph. Nevertheless, we may fix an order for the nodes and have the algorithm always choose the 'smallest' available node. Under these assumptions, EULER$(\mathbf{u})$ is always well defined. Let EULER$(\mathbf{u})$ denote the word of $[\![Q]\!]^n$ obtained from this restricted Eulerian walk. It remains to verify that EULER$(\mathbf{u}) = \mathbf{x}$.

As mentioned in Section 2, an element in $\mathcal{Q}(n; S)$ is an equivalence class $X \subset [\![q]\!]^n$, where $\mathbf{x}, \mathbf{x}' \in X$ implies that $\mathbf{p}(\mathbf{x}; S) = \mathbf{p}(\mathbf{x}'; S)$. Here, we fix the choice of representative for X. As hinted by the previous discussion, we let this representative be EULER $(\mathbf{p}(\mathbf{y}; S))$ for some $\mathbf{y} \in Y$ and observe that this definition is independent of the choice of $\mathbf{y}$. Then with this choice of representatives, the function EULER indeed decodes a profile vector back to its representative codeword.

In summary, we identify the elements in $\mathcal{Q}(n; S)$ with the set of representatives $\{\text{EULER}(\mathbf{u}) : \mathbf{u} \in \mathbf{p}\mathcal{Q}(n; S\}$. Then for any $\mathbf{x} \in \mathcal{Q}(n; S)$, the function EULER decodes $\mathbf{p}(\mathbf{x}; S)$ to $\mathbf{x}$ in linear-time.

A. Practical Methods for Counting ℓ-grams

An interesting feature of the described coding scheme is that one can avoid common problems with DNA sequence assembly by designing codewords that have distinct profile vectors and profiles at sufficiently large distance. However, there are computational challenges associated with counting the number of ℓ-grams and determining the profile vector of an arbitrary word, given that modern high-throughput sequences may produce hundreds of millions of reads. We examine next a number of practical methods for profile counting which represents a crucial step in decoding and address emerging issues via known coding solutions.

In particular, we look at an older technology – sequencing by hybridization (SBH), proposed in [35] – as a means of automated decoding. The idea behind SBH is to build an array of ℓ-grams or *probes*; this array of probes is commonly referred to as a *sequencing chip*. A sample of single stranded DNA to be sequenced is fragmented, labelled with a radioactive or fluorescent material, and then presented to the chip. Each probe in the array hybridizes with its reverse complement, provided the corresponding ℓ-gram is present in the sample. Then an optical detector measures the intensity of hybridization of the labelled DNA and hence infers the number of ℓ-grams present in the sample. The advantage of using SBH for counting ℓ-grams is massive parallelism, and hence increased speed of decoding. Furthermore, SBH allows one to bypass the reading step in sequencing as this is automatically accomplished via hybridization to a proper target.

We first present an analysis of the simplest form of SBH, in which hybridization results may only indicated the presence or absence of certain ℓ-grams. This simple and inexpensive sequencing method may be used to significantly reduce the space of possible profile vectors, and this information may be used to design a more cost efficient and accurate SBH sequencer having fewer probes and more precise probe binding intensity – and hence ℓ-gram counts.

In our discussion, we assume that $S = \llbracket q \rrbracket^\ell$. Furthermore, in our terminology, if $\mathbf{x}$ is the codeword, the channel outputs a subset of $\llbracket q \rrbracket^\ell$ given by $\text{supp}(\mathbf{p}(\mathbf{x}; q, \ell))$, where $\text{supp}(\mathbf{u})$ denotes the set of coordinates $\mathbf{z}$ with $u_{\mathbf{z}} \geq 1$ (see Fig. 4(a)). Then, we can define $d^*_{\text{gram}}(\mathbf{x}, \mathbf{y}; q, \ell) \triangleq |\text{supp}(\mathbf{p}(\mathbf{x}; q, \ell)) \Delta \text{supp}(\mathbf{p}(\mathbf{y}; q, \ell))|$ for any pair of $\mathbf{x}, \mathbf{y} \in \llbracket q \rrbracket^n$. Intuitively, d^*_{gram} measures how dissimilar the sets of ℓ-grams contained in two sequences are.

As before, $(\llbracket q \rrbracket^n, d^*_{\text{gram}})$ forms a pseudometric space and we convert this space into a metric space via an equivalence relation – we say $\mathbf{x} \overset{\ell^*}{\sim} \mathbf{y}$ if and only if $d^*_{\text{gram}}(\mathbf{x}, \mathbf{y}; q, \ell) = 0$. Then, by defining $\mathcal{Q}^*(n; q, \ell) \triangleq \llbracket q \rrbracket^n / \overset{\ell^*}{\sim}$, we obtain a metric space.

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 said to be $(n, d; q, \ell)$-ℓ^*-gram reconstruction code ($*$-GRC). Intuitively, a $*$-GRC with high distance allows for the reconstruction of any codeword sequence via the measurement of a sufficiently large subset of the ℓ-grams. We have the following proposition that is an analogue of Proposition 3.2.

Proposition 8.1. Given an $(n, d; q, \ell)$-$*$-GRC, a set of $n - \ell + 1 - \lfloor (d-1)/2 \rfloor$ ℓ-grams suffices to identify a codeword.

Proof: Let $t = n - \ell + 1 - \lfloor (d-1)/2 \rfloor$. Suppose otherwise that there exists a pair of distinct codewords $\mathbf{x}$ and $\mathbf{y}$ that contain a common set of t ℓ-grams. Then

$$\begin{aligned} d^*_{\text{gram}}(\mathbf{x}, \mathbf{y}; \ell) &= |\text{supp}(\mathbf{p}(\mathbf{x}; q, \ell)) \Delta \text{supp}(\mathbf{p}(\mathbf{y}; q, \ell))| \\ &\leq (n - \ell + 1 - t) + (n - \ell + 1 - t) \\ &= 2 \lfloor (d-1)/2 \rfloor) \leq d - 1 < d, \end{aligned}$$

resulting in a contradiction. □

Determining the maximum size of an $(n, d; q, \ell)$-$*$-GRC turns out to be related to certain well studied combinatorial problems.

Case $d = 1$. The maximum size of an $(n, 1; q, \ell)$-$*$-GRC is given by $|\mathcal{Q}^*(n; q, \ell)|$. Equivalently, this count corresponds to the number of possible sets of ℓ-grams that can be obtained from words of length n. Observe that $|\mathcal{Q}^*(n; q, \ell)| \leq 2^{q^\ell}$ and hence $|\mathcal{Q}^*(n; q, \ell)|$ cannot be a quasipolynomial in n with degree at least one. Therefore, it appears that Ehrhart theory is not applicable in this context. Nevertheless, preliminary investigations of this quantity for $q = 2$ have been performed by Tan and Shallit [36]. In particular, Tan and Shallit proved the following proposition for $n < 2\ell$.

Proposition 8.2 ([36, Corollary 19])**.** For $\ell \leq n < 2\ell$, we have

$$\mathcal{Q}(n; 2, \ell) = 2^n - \sum_{k=1}^{n-\ell+1} \frac{k-1}{k} \sum_{d|k} \mu\left(\frac{k}{d}\right) 2^d,$$

where $\mu(\cdot)$ is the Möbius function defined as

$$\mu(n) = \begin{cases} (-1)^{\omega(n)}, & \text{if } n \text{ is a square-free positive integer;} \\ 0, & \text{otherwise,} \end{cases}$$

and $\omega(n)$ is the number of prime factors of n.

Case $d = 2(n - \ell + 1)$. For the other extreme, we see that the problem is related to edge-disjoint path packings and decompositions of graphs (see [37], [38]). Formally, consider a graph G. A set $\mathcal{C}$ of paths in G is said to be an *edge-disjoint path packing* of G if each edge in G appears in at most one path in $\mathcal{C}$. An edge-disjoint path packing $\mathcal{C}$ of G is an *edge-disjoint path decomposition* of G if each edge in G appears in exactly one path in $\mathcal{C}$. Edge-disjoint cycle packings and decompositions are defined similarly.

Now, an $(n, 2(n - \ell + 1); q, \ell)$-$*$-GRC is equivalent to an edge-disjoint path packing of $D(q, \ell)$, where each path is of length $(n - \ell + 1)$. Furthermore, an edge-disjoint path decomposition of $D(q, \ell)$ into paths of length $n - \ell + 1$ yields an optimal $(n, 2(n - \ell + 1); q, \ell)$-$*$-GRC of size $q^\ell/(n - \ell + 1)$.

Since an edge-disjoint cycle decomposition is also an edge-disjoint path decomposition, we examine next edge-disjoint

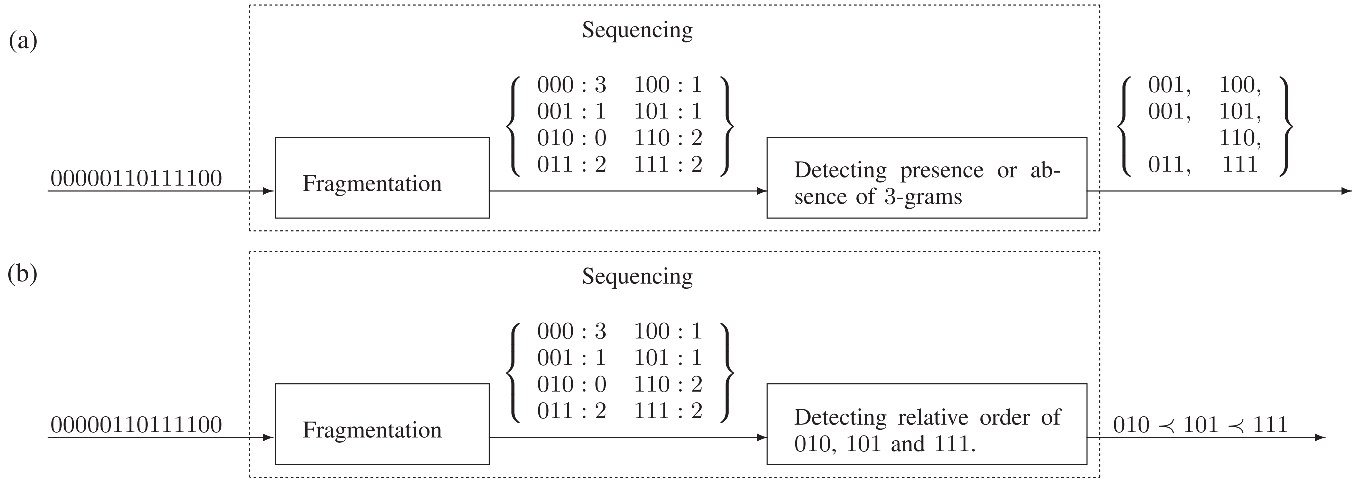


Fig. 4. Sequencing by hybridization. Instead of obtaining the exact count of the ℓ-grams, we obtain auxiliary information on the count: (a) we obtain the set of 3-grams present in 00111011000000; (b) we obtain the relative order of the counts of 010, 101 and 111.

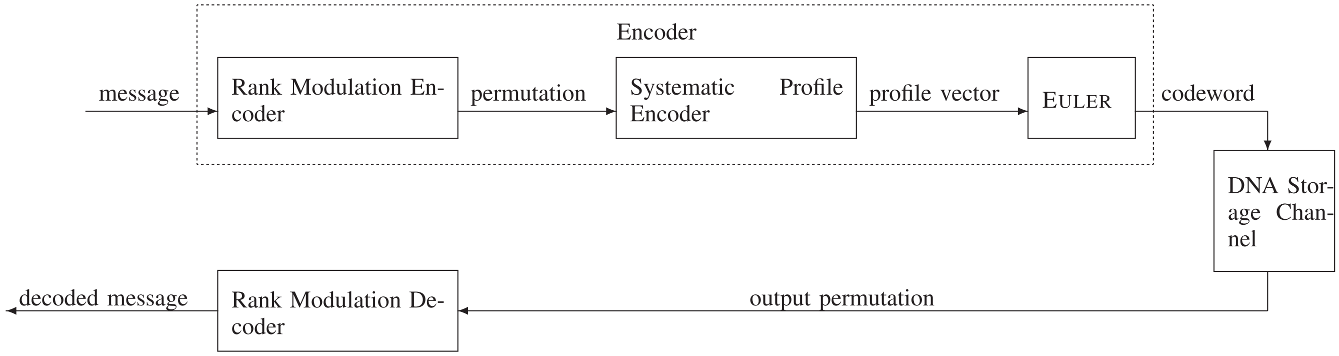


Fig. 5. Encoding messages for a DNA storage channel that outputs the relative order on the counts of particular ℓ-grams.

cycle decomposition of de Bruijn graphs. These combinatorial objects were studied by Cooper and Graham, who proved the following theorem.

Theorem 8.3 ([39, Proposition 2.3, Corollary 2.5])**.**

(i) There exists an edge-disjoint cycle decomposition of $D(q,\ell)$ into q cycles of length $q^{\ell-1}$, for any q and ℓ.

(ii) There exists an edge-disjoint cycle decomposition of $D(r2^{k+1},3)$ into 8^k cycles of length $8r^3$, for any $k \geq 0$ and $r \geq 1$.

Therefore, Theorem 8.3 demonstrates the existence of an optimal $(q^{\ell-1}+\ell-1, 2q^{\ell-1}; q,\ell)$-$*$-GRC of size q and an optimal $(8r^3+2, 16r^3; r2^{k+1}, 3)$-$*$-GRC of size 8^k for any $k \geq 0$ and $r \geq 1$.

B. Decoding Rank Modulation Encoded Profiles

As mentioned earlier, it is difficult to infer accurately the number of ℓ-grams present from the hybridization results. However, we may significantly more accurately determine whether the count of a certain ℓ-gram is greater than the count of another. In other words, we may view the sequencing channel outputs as *rankings* or *orderings* on the q^ℓ ℓ-grams counts or a *permutation* of length q^ℓ reflecting the ℓ-gram counts.

This suggests that we consider codewords whose profile vectors *carry information about order*. More precisely, let $\mathrm{Perm}(N)$ denote the set of permutations over the set $[\![N]\!]$. We consider codewords whose profile vectors belong to $\mathrm{Perm}(N)$ and consider a metric on $\mathrm{Perm}(N)$ that relates to errors resulting from changes in order. The Kendall metric was first proposed by Jiang *et al.* [40] in rank modulation schemes for nonvolatile flash memories and codes in this metric have been studied extensively since (see [41] and the references therein). The Ulam metric was later proposed by Farnoud *et al.* for permutations [42] and multipermutations [43].

Unfortunately, due to the flow conservation equations (1), the profile vector of a q-ary word is unlikely to have distinct entries and hence be a permutation. Nevertheless, we appeal to the systematic encoder provided by Theorem 6.4. We set $m = q^\ell - q^{\ell-1} - 1$. Then, provided n is sufficiently large, there exists a set I of m coordinates that allow us to extend any word $\mathbf{v}$ in $[\![m]\!]^m$ to a profile vector in $\phi_{\rm sys}(\mathbf{v}) \in \mathbf{p}\mathcal{Q}(n;q,\ell)$. In particular, since $\mathrm{Perm}(m) \subseteq [\![m]\!]^m$, *any permutation* $\mathbf{v}$ of length m may be extended to a profile vector in $\phi_{\rm sys}(\mathbf{v}) \in \mathbf{p}\mathcal{Q}(n;q,\ell)$.

This implies that for the design of the sequencing chip, we do not need to have q^ℓ probes for all possible ℓ-grams. Instead, we require only $m = q^\ell - q^{\ell-1} - 1$ probes that correspond to the ℓ-grams in I. Hence, the sequencing channel outputs an ordering on this set of m ℓ-grams (see Fig. 4(b)).

This setup allows us to integrate known rank modulation codes (in any metric) into our coding schemes for DNA storage. In particular, to encode information we perform the following procedure. First, we encode a message into a permutation using a rank modulation encoder. Then the permutation is extended into a profile vector and then mapped by EULER to the profile vector of a q-ary codeword (see Fig. 5 for an illustration).

Example 8.1. Suppose that $S = \llbracket 2 \rrbracket^3$. Hence, we set $m = 3$ and recall the systematic encoder $\phi_{\rm sys}$ described in Example 6.1 that maps $\llbracket 3 \rrbracket^3$ into $\mathbf{p}\mathcal{Q}(14; 2, 3)$. Suppose that $\mathbf{v} = (0, 1, 2) \in \text{Perm}(3)$ belongs to some rank modulation code. Then $\mathbf{u} = \phi_{\rm sys}(\mathbf{v}) = (3, 1, \mathbf{0}, 2, \mathbf{1}, 1, 2, \mathbf{2})$ belongs to $\mathbf{p}\mathcal{Q}(14; 2, 3)$. Finally, EULER maps $\mathbf{u}$ to a codeword $00000110111100 \in \llbracket 2 \rrbracket^{14}$.

Now, if we were to detect the relative order of the 3-grams 010, 101 and 111, we obtain the permutation $(0, 1, 2)$ as desired (see also Fig. 4(b)).

9. Conclusions and Open Problems

We introduced a new coding method tailored for the need of DNA-based storage systems that synthesize DNA strands with potential substitution errors and sequence the strands using shotgun sequencing methods. The synthesis and sequencing methods introduce previously unknown code constraints, as the input to the corresponding DNA storage channel is a sequence, while the output of the DNA storage channel is a collection of (possibly noisy) substrings of the original string, all of predetermined fixed length. The investigated model assumes that only a small bounded number of substrings are in error, and that some substrings may not be observed due to coverage errors. The gist of the approach is to implement error-correction at the level of sequence profiles, where a profile summarizes the number of substrings of each type observed at the channel output. Given that sequence profiles are related to flows in de Bruijn graphs, finding the size of the largest profile code may be cast as a problem of counting lattice points in a rational polytope. To obtain bounds on this count, we used results from Erhart-Macdonald's reciprocity theory. This theory may be seen as broad generalization of the simple result known as Pick's theorem, which allows one to determine the area of a polygon in terms of the number of lattice points in its interior [44].

This work is the first in a line of recent papers on coding for DNA-based storage [45]–[47]. The aforementioned papers addressed the coding challenges associated with different sequencing technologies (i.e., coding for nanopore sequencers [45]), coding problems related to address design for random access DNA-based storage (i.e., mutually uncorrelated and weakly mutually uncorrelated sequences and codes [46]. Note that a special class of this sequences is known as cross-bifix free codes [48]), and codes capable of handling DNA aging which manifests itself through sequence breakage and rearrangement (i.e., codes in the Damerau distance [47]). Nevertheless, many open problems remain, mainly due to the fact that different synthesis and sequencing technologies introduce different types of errors and tend to produce widely different sequence outputs. As an illustrative example, MinION nanopore sequencers produce readouts of current changes corresponding to short consecutive substrings of the sequenced DNA strand, where the current change reflects the chemical structure and composition of the substring. Hence, errors are highly context dependent, and in addition involve deletions and insertions that arise due to undesired shifting of the sequence within the nanopore. For more details on this and other open coding problems for DNA-based storage, as well as an overview of modern synthesis and sequencing technologies, the interested reader is referred to the overview paper [9].

Appendix A
Table of Definitions and Notation

Symbol	Meaning	Remarks
n	Length of sequence or word	Sec. 2
q	Size of the alphabet	Sec. 2
$\llbracket q \rrbracket$	$\{0, 1, \ldots, q-1\}$	Sec. 2
ℓ-gram	Substring of length ℓ, a subsequence corresponding to elements with ℓ consecutive indices, or a read	Sec. 2
$\mathbf{x}$	Input sequence	Sec. 1
$\widetilde{\mathbf{x}}$	String obtained through biochemical synthesis of the string $\mathbf{x}$	Sec. 1, Def. 2.1
$\widetilde{\mathbf{p}}(\mathbf{x})$	Multiplicity vector of $\widetilde{\mathbf{x}}$	Sec. 1, Def. 2.1
$\widehat{\mathcal{L}}(\mathbf{x})$	Unordered subset of substrings of $\widetilde{\mathbf{x}}$	Sec. 1
$\widehat{\mathbf{p}}(\mathbf{x})$	Output profile vector of $\mathbf{x}$, or multiplicity vector of $\widehat{\mathcal{L}}(\mathbf{x})$	Sec. 1, Def. 2.1
$\mathbf{p}(\mathbf{x}; q, \ell)$	ℓ-gram profile vector of $\mathbf{x}$	Sec. 2
S	set of ℓ-grams	Sec. 2
$S(q, \ell; q^*, [w_1, w_2])$	$\{\mathbf{x} \in \llbracket q \rrbracket^\ell : \text{wt}(\mathbf{x}, q^*) \in [w_1, w_2]\}$, where wt denotes the Hamming weight of the underlying argument and $1 \le q^* \le q-1$	Sec. 2
$\mathbf{p}(\mathbf{x}; S)$	profile vector of $\mathbf{x}$, where all ℓ-grams of $\mathbf{x}$ belong to S	Sec. 2
$(\llbracket q \rrbracket^n; S)$	q-ary words of length n whose ℓ-grams belong to S	Sec. 3
$d_{\rm gram}(\mathbf{x}, \mathbf{y}; S)$	The ℓ-gram distance is the asymmetric distance between the profile vectors of $\mathbf{x}$ and $\mathbf{y}$, where $\mathbf{x}, \mathbf{y} \in (\llbracket q \rrbracket^n; S)$	Sec. 3
$\mathcal{Q}(n; S)$	The set of equivalence classes defined by the relation $\overset{d_{\rm gram}}{\sim}$	Sec. 3
$\mathbf{p}\mathcal{Q}(n; S)$	The set of profile vectors of words in $\mathcal{Q}(n; S)$	Sec. 3
$(n, d; S)$-GRC	$(n, d; S)$-ℓ-gram reconstruction code As we identify words in $\mathcal{Q}(n; S)$ with their corresponding profile vectors in $\mathbf{p}\mathcal{Q}(n; S)$, we slightly abuse notation by referring to GRCs as subsets of $\mathcal{Q}(n; S)$ and $\mathbf{p}\mathcal{Q}(n; S)$, interchangeably.	Sec. 3
$D(S)$	restricted de Brujin graph corresponding to S	Sec. 4
$\bar{\mathcal{Q}}(n; S)$	set of closed words in $\mathcal{Q}(n; S)$	Sec. 4
$\mathbf{p}\bar{\mathcal{Q}}(n; S)$	set of profile vectors corresponding to words in $\mathcal{Q}(n; S)$	Sec. 4

Appendix B
Eulerian Property of Certain Restricted De Bruijn Digraphs

In this section, we provide a detailed proof of Proposition 4.2. Specifically, for q, ℓ, $1 \le q^* \le q-1$ and $1 \le w_1 <$

$w_2 \leq \ell$, we demonstrate that the digraph $D(q, \ell; q^*, [w_1, w_2])$ is Eulerian. Our analysis follows that of Ruskey *et al.* [23].

Recall that the arc set of $D(q, \ell; q^*, [w_1, w_2])$ is given by $S = S(q, \ell; q^*, [w_1, w_2])$, while the node set is given by $V(S) = S(q, \ell - 1; q^*, [w_1 - 1, w_2])$, which we denote by V for short. In addition, we introduce the following subsets of $[\![q]\!]$. For a node $\mathbf{z}$ in V, let $\mathrm{Pref}(\mathbf{z})$ be the set of symbols in $[\![q]\!]$ that when prepended to $\mathbf{z}$ results in an arc in S. Similarly, let $\mathrm{Suff}(\mathbf{z})$ be the set of symbols in $[\![q]\!]$ that when appended to $\mathbf{z}$ result in an arc in S. Hence, $\{\sigma\mathbf{z} : \sigma \in \mathrm{Pref}(\mathbf{z})\}$ and $\{\mathbf{z}\sigma : \sigma \in \mathrm{Suff}(\mathbf{z})\}$ are the respective sets of incoming and outgoing arcs for the node $\mathbf{z}$.

Lemma 9.1. Every node of $D(q, \ell; q^*, [w_1, w_2])$ has the same number of incoming and outgoing arcs.

Proof: Let $\mathbf{z}$ belong to V. Observe that for all $s \in [\![q]\!]$, $s\mathbf{z} \in S$ if and only if $\mathbf{z}s \in S$. Hence, $\mathrm{Pref}(\mathbf{z}) = \mathrm{Suff}(\mathbf{z})$ and the lemma follows. □

It remains to show that $D(q, \ell; q^*, [w_1, w_2])$ is strongly connected. We do it via the following sequence of lemmas.

Lemma 9.2. Let $\mathbf{z}, \mathbf{z}'$ belong to V and have the property that they differ in exactly one coordinate. Then there exists a path from $\mathbf{z}$ to $\mathbf{z}'$.

Proof: Observe the following characterization of $\mathrm{Pref}(\mathbf{z}) = \mathrm{Suff}(\mathbf{z})$:

$$\begin{aligned}\mathrm{Pref}(\mathbf{z}) &= \mathrm{Suff}(\mathbf{z}) \\ &= \begin{cases} [q - q^*, q - 1], & \text{if } \mathrm{wt}(\mathbf{z}; q^*) = w_1 - 1; \\ [\![q^*]\!], & \text{if } \mathrm{wt}(\mathbf{z}; q^*) = w_2; \\ [\![q]\!], & \text{otherwise.} \end{cases}\end{aligned}$$

Then $\mathrm{Suff}(\mathbf{z}) \cap \mathrm{Pref}(\mathbf{z}')$ is empty only if $\mathrm{wt}(\mathbf{z}; q^*) = w_1 - 1$ and $\mathrm{wt}(\mathbf{z}'; q^*) = w_2$ or vice versa. Either way, $\mathbf{z}$ and $\mathbf{z}'$ differ in at least two coordinates, which contradicts the starting assumption.

Hence, $\mathrm{Suff}(\mathbf{z}) \cap \mathrm{Pref}(\mathbf{z}')$ is always nonempty. To complete the proof, let $s \in \mathrm{Suff}(\mathbf{z}) \cap \mathrm{Pref}(\mathbf{z}')$. Then, the path corresponding to $\mathbf{z}s\mathbf{z}'$ is the desired path. (Note that each ℓ-gram appearing in $\mathbf{z}s\mathbf{z}'$ has weight equal to either $\mathrm{wt}(\mathbf{z}s)$ or $\mathrm{wt}(s\mathbf{z}')$; in particular, each such ℓ-gram lies in S.) □

Therefore, to construct a path between any two given nodes $\mathbf{z}$ and $\mathbf{z}'$, it suffices to demonstrate a sequence of nodes such that consecutive nodes differ in only one position.

Lemma 9.3. For any $\mathbf{z}, \mathbf{z}' \in V$, there is a sequence of nodes $\mathbf{z} = \mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_t = \mathbf{z}'$ such that $\mathbf{z}_j$ and $\mathbf{z}_{j+1}$ differ in exactly one position for $j \in [\![t]\!]$.

Proof: Let $\mathbf{z}' = \sigma_1\sigma_2\cdots\sigma_{\ell-1}$. We construct the sequence of nodes inductively. Suppose that for some j, $\mathbf{z}_j = \sigma_1\sigma_2\cdots\sigma_i\tau_{i+1}\cdots\tau_{\ell-1}$, with $\tau_{i+1} \neq \sigma_{i+1}$. Our objective is to construct a sequence of nodes with consecutive nodes differing in one position, terminating at some node $\mathbf{z}_{j'}$ with $\mathbf{z}_{j'} = \sigma_1\sigma_2\cdots\sigma_i\sigma_{i+1}\tau'_{i+2}\cdots\tau'_{\ell-1}$ for some $\tau'_{i+1}, \tau'_{i+2}, \ldots, \tau'_{\ell-1}$. Hence, by repeating this procedure, we obtain the desired sequence of nodes that terminates at $\mathbf{z}'$.

Since $\mathbf{z}_j \in V$, we have $\mathrm{wt}(\mathbf{z}_j; q^*) \in [w_1 - 1, w_2]$. As such, we consider three possibilities to extend the sequence:

(i) When $w_1 - 1 < \mathrm{wt}(\mathbf{z}_j; q^*) < w_2$, we may simply change τ_{i+1} to σ_{i+1} and make no other changes, since the word $\mathbf{z}_{j+1}$ produced this way still satisfies $\mathrm{wt}(\mathbf{z}_{j+1}) \in [w_1 - 1, w_2]$ and is therefore a node.

(ii) When $\mathrm{wt}(\mathbf{z}_j; q^*) = w_1 - 1$, $\tau_{i+1} \in [q - q^*, q - 1]$ and $\sigma_{i+1} \notin [q - q^*, q - 1]$, there exists some τ_k in $\mathbf{z}_j$ that does not belong to $[q - q^*, q - 1]$. Otherwise, $\mathrm{wt}(\sigma_1\cdots\sigma_i; q^*) = w_1 - \ell + i$ and so $\mathrm{wt}(\sigma_1\cdots\sigma_{i+1}; q^*) = w_1 - \ell + i$. Then, $\mathrm{wt}(\mathbf{z}'; q^*) \leq w_1 - 2$, contradicting the fact that $\mathbf{z}' \in V$. Therefore, we have the sequence of nodes

$$\begin{aligned}\mathbf{z}_j &= \sigma_1\cdots\sigma_i\tau_{i+1}\tau_{i+2}\cdots\tau_k\cdots\tau_{\ell-1}, \\ \mathbf{z}_{j+1} &= \sigma_1\cdots\sigma_i\tau_{i+1}\tau_{i+2}\cdots(q-1)\cdots\tau_{\ell-1}, \\ \mathbf{z}_{j+2} &= \sigma_1\cdots\sigma_i\sigma_{i+1}\tau_{i+2}\cdots(q-1)\cdots\tau_{\ell-1}.\end{aligned}$$

(iii) When $\mathrm{wt}(\mathbf{z}_j; q^*) = w_2$, $\tau_{i+1} \notin [q - q^*, q - 1]$ and $\sigma_{i+1} \in [q - q^*, q - 1]$, then there exists some τ_k in $\mathbf{z}_j$ that belongs to $[q - q^*, q - 1]$. Otherwise, $\mathrm{wt}(\sigma_1\cdots\sigma_i; q^*) = w_2$ and so $\mathrm{wt}(\mathbf{z}'; q^*) \geq \mathrm{wt}(\sigma_1\cdots\sigma_{i+1}; q^*) = w_2 + 1$, contradicting the fact that $\mathbf{z}' \in V$. Therefore, we have the sequence of nodes

$$\begin{aligned}\mathbf{z}_j &= \sigma_1\cdots\sigma_i\tau_{i+1}\tau_{i+2}\cdots\tau_k\cdots\tau_{\ell-1}, \\ \mathbf{z}_{j+1} &= \sigma_1\cdots\sigma_i\tau_{i+1}\tau_{i+2}\cdots 0\cdots\tau_{\ell-1}, \\ \mathbf{z}_{j+2} &= \sigma_1\cdots\sigma_i\sigma_{i+1}\tau_{i+2}\cdots 0\cdots\tau_{\ell-1}.\end{aligned}$$

□

Consequently, $D(q, \ell; q^*, [w_1, w_2])$ is strongly connected. Together with Lemma 9.1, this result establishes that $D(q, \ell; q^*, [w_1, w_2])$ is Eulerian.

Appendix C
Proof of Corollary 5.6

We provide next a detailed proof of Corollary 5.6. Specifically, we demonstrate Proposition 9.4 from which the corollary follows directly. For the case that $S = [\![q]\!]^\ell$, Jacquet *et al.* established a similar result by analyzing a sum of multinomial coefficients. This type of analysis appears to be to complex for a general choice of S.

Proposition 9.4. Suppose that $D(S)$ is strongly connected and that it contains loops. Let $t = n - \ell + 1$, $D = |S| - |V(S)|$ and let the lattice point enumerator of $\mathcal{P}(S)$ be $L_{\mathcal{P}(S)}(t) = c_D(t)t^D + O(t^{D-1})$. Then, $c_D(t)$ is constant.

To prove this proposition, we use the following straightforward lemma.

Lemma 9.5. Suppose that $D(S)$ is strongly connected and that it contains loops. For all t, we have $L_{\mathcal{P}(S)}(t + 1) \geq L_{\mathcal{P}(S)}(t)$.

Proof: It suffices to show that there is an injection from $\mathcal{LP}_{\geq 0}(n; S)$ to $\mathcal{LP}_{\geq 0}(n+1; S)$. Suppose that $\mathbf{u} \in \mathcal{LP}_{\geq 0}(n; S)$, so that $\mathbf{A}(S)\mathbf{u} = t\mathbf{b}$. Fix a loop in $D(S)$ and consider the vector $\boldsymbol{\chi}(\mathbf{z})$, where $\mathbf{z}$ is the arc corresponding to the loop. Then, $\mathbf{A}(S)\boldsymbol{\chi}(\mathbf{z}) = \mathbf{b}$ and $\mathbf{A}(S)(\mathbf{u} + \boldsymbol{\chi}(\mathbf{z})) = (t + 1)\mathbf{b}$. So, the map $\mathbf{u} \mapsto \mathbf{u} + \boldsymbol{\chi}(\mathbf{z})$ is an injection from $\mathcal{LP}_{\geq 0}(n; S)$ to $\mathcal{LP}_{\geq 0}(n + 1; S)$. □

Proof: [Proof of Proposition 9.4] Lemma 9.5 demonstrates that $L_{\mathcal{P}(S)}$ is a monotonically increasing function. Intuitively, this implies that the coefficient of its dominating term $c_D(t)$ cannot be periodic with period greater than 1. We prove this claim formally in what follows.

Suppose that c_D is not constant and that it has period τ. Hence, there exists $t_a \not\equiv t_b \bmod \tau$ such that $c_D(t_a) = a_D$, $c_D(t_b) = b_D$ and $a_D < b_D$. Furthermore, define $a_i = c_i(t_a)$ and $b_i = c_i(t_b)$ for $0 \le i \le D-1$, and consider the polynomial $\sum_{i=0}^{D} b_i t^i - a_i(t+\tau)^i$. By construction, this polynomial has degree D and a positive leading coefficient. Hence, we can choose $t_1 \equiv t_a \bmod \tau$ and $t_2 \equiv t_b \bmod \tau$ so that $t_1 \le t_2 \le t_1 + \tau$ and $\sum_{i=0}^{D} b_i t_2^i - a_i(t_1+\tau)^i > 0$. Consequently,

$$\begin{aligned} L_{\mathcal{P}(S)}(t_1+\tau) &= \sum_{i=0}^{D} c_i(t_1+\tau)(t_1+\tau)^i \\ &= \sum_{i=0}^{D} a_i(t_1+\tau)^i \\ &< \sum_{i=0}^{D} b_i t_2^i = L_{\mathcal{P}(S)}(t_2), \end{aligned}$$

contradicting the monotonicity of $L_{\mathcal{P}(S)}$. □

Appendix D
Properties of the Polytope $\mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$

We derive properties of the polytope $\mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$ described in Section 6-A. In particular, under the assumption that $D(S)$ is strongly connected and $\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \text{Null}_{>\mathbf{0}}\mathbf{B}(D(S))$ is nonempty, we demonstrate the following:

(C1) The dimension of the polytope $\mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$ is $|S| - |V(S)|$;

(C2) The interior of the polytope is given by $\{\mathbf{u} \in \mathbb{R}^{|S|+d} : \mathbf{A}(\mathbf{H}, S)\mathbf{u} = \mathbf{b}, \mathbf{u} > \mathbf{0}\}$;

(C3) The vertex set of the polytope is given by

$$\left\{ \left(\frac{\boldsymbol{\chi}(C)}{|C|}, \frac{\mathbf{H}\boldsymbol{\chi}(C)}{p|C|} \right) : C \text{ is a cycle in } D(S) \right\}.$$

Since $\mathcal{C}(\mathbf{H}, \mathbf{0}) \cap \text{Null}_{>\mathbf{0}}\mathbf{B}(D(S))$ is nonempty, let $\mathbf{u}_0$ belong to this intersection. Then $\mathbf{H}\mathbf{u}_0 \equiv \mathbf{0} \bmod p$, that is, $\mathbf{H}\mathbf{u}_0 = p\boldsymbol{\beta}$ for some $\boldsymbol{\beta} > 0$. Let $\mu = \mathbf{1}\mathbf{u}_0$. If we set $\mathbf{u} = \frac{1}{\mu}(\mathbf{u}_0, \boldsymbol{\beta})$, then $\mathbf{A}(\mathbf{H}, S)\mathbf{u} = \mathbf{b}$, with $\mathbf{u} > 0$.

Observe that the block structure of $\mathbf{A}(\mathbf{H}, S)$ implies that it has rank $|V(S)| + d$. Hence, the nullity of $\mathbf{A}(\mathbf{H}, S)$ is $|S| - |V(S)|$. As before, let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{|S|-|V(S)|}$ be linearly independent vectors that span the null space of $\mathbf{A}(\mathbf{H}, S)$. Since $\mathbf{u}$ has strictly positive entries, we can find ϵ small enough so that $\mathbf{u} + \epsilon\mathbf{u}_i$ belongs to $\mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$ for all $i \in [|S| - |V(S)|]$. Therefore, $\{\mathbf{u}, \mathbf{u} + \epsilon\mathbf{u}_1, \mathbf{u} + \epsilon\mathbf{u}_2, \ldots, \mathbf{u} + \epsilon\mathbf{u}_{|S|-|V(S)|}\}$ is a set of $|S| - |V(S)| + 1$ affinely independent points in $\mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$. This proves claim (C1).

For the interior of $\mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$, first consider $\mathbf{u}' > \mathbf{0}$ such that $\mathbf{A}(\mathbf{H}, S)\mathbf{u}' = \mathbf{b}$. For any $\mathbf{u}'' \in \mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$, we have $\mathbf{A}(\mathbf{H}, S)\mathbf{u}'' = \mathbf{b}$ and hence, $\mathbf{A}(\mathbf{H}, S)(\mathbf{u}' - \mathbf{u}'') = \mathbf{0}$. Since $\mathbf{u}'$ has strictly positive entries, we choose ϵ small enough so that $\mathbf{u}' + \epsilon(\mathbf{u}' - \mathbf{u}'') \ge \mathbf{0}$. Therefore, $\mathbf{u}' + \epsilon(\mathbf{u}' - \mathbf{u}'')$ belongs to $\mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$ and $\mathbf{u}'$ belongs to the interior of $\mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$.

Conversely, let $\mathbf{u}' \in \mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$ with $u'_j = 0$ for some coordinate j. Let $\mathbf{u}$ be as defined earlier, where $\mathbf{u} \in \mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$ with $\mathbf{u} > \mathbf{0}$. Hence, for all $\epsilon > 0$, the jth coordinate of $\mathbf{u}' + \epsilon(\mathbf{u}' - \mathbf{u})$ is given by $-\epsilon u_j$, which is always negative. In other words, $\mathbf{u}'$ does not belong to interior of $\mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$. This characterizes the interior as described in claim (C2).

For the vertex set, observe that

$$\left\{ \left(\frac{\boldsymbol{\chi}(C)}{|C|}, \frac{\mathbf{H}\boldsymbol{\chi}(C)}{p|C|} \right) : C \text{ is a cycle in } D(S) \right\} \subseteq \mathcal{P}_{\text{GRC}}(\mathbf{H}, S).$$

Let $\mathbf{v} \in \mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$ and suppose that $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$ is a vertex. Since $\mathbf{v} \in \mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$, we have $\mathbf{v}_2 = \frac{1}{p}\mathbf{H}\mathbf{v}_1$ and $\mathbf{B}(D(S))\mathbf{v}_1 = \mathbf{0}$. Proceeding as in the proof of Lemma 5.5, we conclude that $\mathbf{v}_1 = \boldsymbol{\chi}(C)/|C|$, for some cycle in $D(S)$ and hence, $\mathbf{v} = \left(\frac{\boldsymbol{\chi}(C)}{|C|}, \frac{\mathbf{H}\boldsymbol{\chi}(C)}{p|C|} \right)$.

Conversely, we show that for any cycle C in $D(S)$, $\left(\frac{\boldsymbol{\chi}(C)}{|C|}, \frac{\mathbf{H}\boldsymbol{\chi}(C)}{p|C|} \right)$ cannot be expressed as a convex combination of other points in $\mathcal{P}_{\text{GRC}}(\mathbf{H}, S)$. Suppose otherwise. Then we consider the first $|S|$ coordinates and we proceed as in the proof of Lemma 5.5 to yield a contradiction. This completes the proof of claim (C3).

Appendix E
Relabelling of Nodes in Proof of Theorem 6.4

In this section, we demonstrate the existence of a cyclic relabelling of nodes that is necessary for the proof of Theorem 6.4. In particular, we prove the following lemma.

Lemma 9.6. Let υ be a positive integer, and $r_1, r_2, \ldots, r_\upsilon$ be υ real values such that $\sum_{i=1}^{\upsilon} r_i = 0$. For convenience, we let $r_{\upsilon+i} = r_i$ for $1 \le i \le \upsilon - 1$. Then there exists $1 \le J \le \upsilon$ such that $\sum_{i=0}^{j} r_{J+i} \ge 0$ for all $0 \le j \le \upsilon - 1$.

Proof: For $1 \le j \le 2\upsilon - 1$, let $R_j = \sum_{i=0}^{j} r_i$ and observe that $R_\upsilon = 0$. Let J be such that $R_J = \min\{R_j : 1 \le j \le 2\upsilon - 1\}$. Since $R_\upsilon = 0$, we have $R_{i+\upsilon} = R_i$ for all $1 \le i \le \upsilon - 1$ and hence, we may assume $1 \le J \le \upsilon$.

Next, we claim that J is the desired index. Indeed, for all $0 \le j \le \upsilon - 1$, observe that

$$\sum_{i=0}^{j} r_{J+i} = R_{J+j} - R_J \ge 0,$$

where the inequality follows from the minimality of R_J. □

References

[1] H. M. Kiah, G. J. Puleo, and O. Milenkovic, “Coding for DNA storage channels,” in *IEEE Inform. Theory Workshop*, Jerusalem, 2015, pp. 1–5.

[2] ——, “Codes for DNA sequence profiles,” in *Proc. IEEE Intl. Symp. Inform. Theory*, Hong Kong, 2015, pp. 841–818.

[3] 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.

[4] 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.

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

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

[7] 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.

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

[9] 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.

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

[11] 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.

[12] W. Wan, L. Lulu, Q. Xu, Z. Wang, Y. Yao, R. Wang, J. Zhang, H. Liu, X. Gao, and J. Hong, "Error removal in microchip-synthesized dna using immobilized muts," *Nucleic acids research*, p. gku405, 2014.

[13] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu, "A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers," *BMC genomics*, vol. 13, no. 1, p. 341, 2012.

[14] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe, "Characterizing and measuring bias in sequence data," *Genome Biol*, vol. 14, no. 5, p. R51, 2013.

[15] 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.

[16] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii, "Base-stacking and base-pairing contributions into thermal stability of the dna double helix," *Nucleic acids research*, vol. 34, no. 2, pp. 564–574, 2006.

[17] 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.

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

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

[20] P. A. Pevzner, "DNA physical mapping and alternating Eulerian cycles in colored graphs," *Algorithmica*, vol. 13, no. 1-2, pp. 77–105, 1995.

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

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

[23] 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.

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

[25] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network flows: theory, algorithms, and applications*. Prentice Hall, 1993.

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

[27] R. P. Stanley, *Enumerative combinatorics*. Cambridge university press, 2011, vol. 1.

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

[29] 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.

[30] P. L. Čebyšev, *Mémoire sur les nombres premiers*. Académie Impériale des Sciences, 1850.

[31] M. Agrawal, N. Kayal, and N. Saxena, "Primes is in p," *Annals of mathematics*, pp. 781–793, 2004.

[32] 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.

[33] 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.

[34] C. Hierholzer, "Über die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren," *Mathematische Annalen*, vol. 6, no. 1, pp. 30–32, 1873.

[35] P. A. Pevzner and R. J. Lipshutz, "Towards DNA sequencing chips," in *Mathematical Foundations of Computer Science 1994*. Springer, 1994, pp. 143–158.

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

[37] K. Heinrich, "Path decomposition," *Le Matematiche*, vol. 47, no. 2, pp. 241–258, 1993.

[38] D. Bryant and S. El-Zanati, *Graph decompositions*, 2nd ed. Chapman & Hall/CRC, 2007, ch. VI.24, pp. 477–486.

[39] J. N. Cooper and R. L. Graham, "Generalized de Bruijn cycles," *Annals of Combinatorics*, vol. 8, no. 1, pp. 13–25, 2004.

[40] A. Jiang, R. Mateescu, M. Schwartz, and J. Bruck, "Rank modulation for flash memories," *IEEE Trans. Inform. Theory*, vol. 55, no. 6, pp. 2659–2673, 2009.

[41] A. Barg and A. Mazumdar, "Codes in permutations and error correction for rank modulation," *IEEE Trans. Inform. Theory*, vol. 56, no. 7, pp. 3158–3165, 2010.

[42] F. Farnoud, V. Skachek, and O. Milenkovic, "Error-correction in flash memories via codes in the Ulam metric," *IEEE Trans. Inform. Theory*, vol. 59, no. 5, pp. 3003–3020, 2013.

[43] F. Farnoud and O. Milenkovic, "Multipermutation codes in the Ulam metric for nonvolatile memories," *Selected Areas in Communications, IEEE Journal on*, vol. 32, no. 5, pp. 919–932, 2014.

[44] G. Pick, "Geometrisches zur zahlenlehre," *Sitzenber. Lotos (Prague)*, vol. 19, pp. 311–319, 1899.

[45] R. Gabrys, H. M. Kiah, and O. Milenkovic, "Asymmetric lee distance codes for dna-based storage," in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 909–913.

[46] S. Yazdi, H. M. Kiah, and O. Milenkovic, "Weakly mutually uncorrelated codes," *arXiv preprint arXiv:1601.08176*, 2016.

[47] R. Gabrys, E. Yaakobi, and O. Milenkovic, "Codes in the damerau distance for dna storage," *arXiv preprint arXiv:1601.06885*, 2016.

[48] D. Bajic and T. Loncar-Turukalo, "A simple suboptimal construction of cross-bifix-free codes," *Cryptography and Communications*, vol. 6, no. 1, pp. 27–37, 2014.

Han Mao Kiah received his Ph.D. degree in mathematics from the Nanyang Technological University, Singapore in 2014. From 2014 to 2015, he was a Postdoctoral Research Associate at the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. Currently, he is a lecturer at the School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore. His research interests include combinatorial design theory, coding theory, and enumerative combinatorics.

Gregory J. Puleo received his Ph.D. in Mathematics from the University of Illinois at Urbana-Champaign in 2014. He is currently a Postdoctoral Research Associate at the Coordinated Science Lab at UIUC. His research interests include extremal graph theory and combinatorial games.

Olgica Milenkovic (M'04, SM'12) received her Ph.D. in Electrical Engineering from the University of Michigan, Ann Arbor and is currently a professor in the Electrical Engineering Department of University of Illinois. She is a recipient of the NSF Career Award, the DARPA Young Faculty Award, and the Dean's Excellence in Research Award. In 2012 she was named a Center for Advanced Studies (CAS) Associate, and in 2013 she became a Willet scholar. In 2015, she was named Distinguished Lecturer of the IEEE Information Theory Society. She served on the editorial board for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE TRANSACTIONS ON INFORMATION THEORY. Her research interests are in bioinformatics, coding theory, compressive sensing and social sciences.