

Constrained Codes over Degenerate-Base Alphabets via Compression of Adjacency Matrices

Dang Hai Dang* and Sunghwan Kim†

*Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Republic of Korea

†Department of Artificial Intelligence and Information Technology, Sejong University, Seoul 05006, Republic of Korea

Emails: haidangel29@gmail.com, ks@sejong.ac.kr

Abstract—Degenerate bases significantly enhance the information capacity of DNA-based storage systems but necessitate strict run-length-limited (RLL) constraints to prevent homopolymer-induced synthesis and sequencing errors. Existing constrained code constructions for these alphabets either rely on numerical capacity estimates or require constraint-specific auxiliary computations that are impractical for RLL constraints over degenerate-base alphabets due to the exponential growth in forbidden substrings. To address these challenges, this paper introduces a compressed matrix construction based on the $\ell \times \ell$ run-extending auxiliary matrix and the $\ell \times \ell$ run-reset auxiliary matrix for constrained length ℓ . The construction yields minimal-degree closed-form constrained-capacity formulations for any degenerate-base alphabet and provides an explicit $O(1)$ -time criterion to identify the capacity-maximizing alphabet without solving capacity roots numerically.

Index Terms—DNA storage, degenerate-base alphabets, RLL constraint, compressed matrix, constrained capacity.

I. INTRODUCTION

DNA-based data storage has received increasing attention as a potential solution for long-term data archiving, since DNA can remain stable over extended periods, store information at very high physical densities, and be reliably read using widely available sequencing technologies [1]–[4]. In conventional DNA storage applications, data are encoded using the four canonical nucleotide bases as discrete letters, forming the standard DNA alphabet. Degenerate bases increase information density by allowing one letter to represent a mixture of the four canonical nucleotide bases [5], [6].

The practical implementation of DNA storage is challenged by errors during synthesis and sequencing. A primary cause of such errors is the formation of long homopolymer runs of nucleotide bases [7]–[9]. For constrained length ℓ , a standard solution is to impose the ℓ -RLL constraint, which prevents any nucleotide base from appearing more than ℓ consecutive times.

For a degenerate-base alphabet of size κ , the comprehensive FSTD of Nguyen *et al.* [10] enumerates all length- ℓ substrings. Its adjacency matrix is used to compute constrained capacity from the spectral radius [11]. Therefore, we obtain a comprehensive adjacency matrix with κ^ℓ states and characteristic polynomial degree κ^ℓ , both exponential in ℓ . A direct closed-form constrained-capacity formula would require factorizing this exponential-degree characteristic polynomial. The Goulden–Jackson cluster theorem [12] can compute a

closed-form generating function from overlaps of forbidden substrings. However, for degenerate-base alphabets, this approach still requires exponentially many forbidden substrings and the factorization of the induced exponential-degree generating function.

In contrast, our goal is to obtain closed-form constrained-capacity formulas specifically for the ℓ -RLL constraint over degenerate-base alphabets. Instead of characterizing all forbidden substrings, we characterize the constraint by the run-extension and run-reset auxiliary $\ell \times \ell$ block matrices (see Construction 1) and compress the induced matrix by colour-refinement (see Construction 2). Conceptually, this has the same purpose as adjacency matrix compression [13], [14], namely constructing an alternative smaller adjacency matrix while preserving the spectral radius and therefore the constrained capacity.

Accordingly, this paper constructs a compressed matrix for any degenerate-base alphabet (see Theorem 1), derives a minimal-degree closed-form constrained-capacity polynomial for the six-letter degenerate-base alphabet (see Theorem 2), and proves the optimality of that alphabet through a recurrence comparison rather than numerical capacity estimates (see Theorem 3).

II. PRELIMINARIES

Definition 1 (ℓ -RLL constraint for the standard DNA alphabet): Let $\Sigma_4 = \{A, C, G, T\}$ be the standard DNA alphabet, with constrained length $\ell \in \mathbb{N}$. A sequence $s = (s_1, \dots, s_n) \in \Sigma_4^n$ satisfies the ℓ -RLL constraint if no letter appears consecutively more than ℓ times. Equivalently, the constraint is violated if $s_m = s_{m+1} = \dots = s_{m+\ell}$ for some $1 \leq m \leq n - \ell$.

Definition 2 (Degenerate-base alphabet): A degenerate-base alphabet Σ_κ is an ordered set of κ letters, where each letter $s \in \Sigma_\kappa$ represents the non-empty support set $\text{supp}(s) \subseteq \Sigma_4$. The order of letters in this alphabet is lexicographic: letters containing A first, followed by C, followed by G, and finally T.

For instance, degenerate-base alphabet $\Sigma_6^+ = \{A, A|C, C, G, G|T, T\}$ contains standard letters A, C, G, T and two degenerate-base letters A|C, G|T, with $\text{supp}(A) = \{A\}$ and $\text{supp}(A|C) = \{A, C\}$.

Definition 3 (ℓ -RLL constraint for degenerate-base alphabets): For constrained length ℓ , a sequence $(s_1, \dots, s_n) \in \Sigma_\kappa^n$ violates the ℓ -RLL constraint if there exist $x \in \Sigma_4$ and an

index m with $1 \leq m \leq n - \ell$ such that $x \in \text{supp}(s_q)$ for every $q \in \{m, m+1, \dots, m+\ell\}$.

Definition 4 (Support-intersection set): For each letter $s \in \Sigma_\kappa$, its support-intersection set is $\mathcal{N}(s) = \{t \in \Sigma_\kappa \mid \text{supp}(s) \cap \text{supp}(t) \neq \emptyset\}$.

III. COMPRESSED MATRIX CONSTRUCTION FOR THE RLL CONSTRAINT

For a degenerate-base alphabet Σ_κ with constrained length ℓ , let $\mathbf{D}(\ell; \Sigma_\kappa)$ be the comprehensive adjacency matrix of the FSTD. Its states are all length- ℓ substrings over Σ_κ , so its size is $\kappa^\ell \times \kappa^\ell$. A transition from $\mathbf{x} = x_1 \cdots x_\ell$ to $\mathbf{y} = y_1 \cdots y_\ell$ is allowed when $x_2 \cdots x_\ell = y_1 \cdots y_{\ell-1}$ and the concatenated length- $(\ell+1)$ substring $x_1 \cdots x_\ell y_\ell$ satisfies the ℓ -RLL constraint. The comprehensive adjacency matrix of the FSTD is not suitable for closed-form derivation, since the characteristic polynomial degree is κ^ℓ .

In this section, we construct an adjacency matrix $\mathbf{A}$ as an alternative to the comprehensive exponential-size FSTD obtained by enumerating all length- ℓ substrings. The construction uses the set $\Sigma_\kappa \times \{1, \dots, \ell\}$ of size $\kappa\ell$, where each element (s, r) consists of a degenerate-base letter $s \in \Sigma_\kappa$ and a run-length r . The auxiliary $\ell \times \ell$ block matrices in Construction 1 construct $\mathbf{A}$, and the construction preserves the constrained capacity of the comprehensive FSTD, namely $\rho(\mathbf{D}(\ell; \Sigma_\kappa)) = \rho(\mathbf{A})$.

Construction 1 (FSTD adjacency matrix from auxiliary $\ell \times \ell$ block matrices): The ℓ -RLL constraint is represented by the $\ell \times \ell$ run-extending auxiliary block matrix

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

and the $\ell \times \ell$ run-reset auxiliary block matrix

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}.$$

The matrix $\mathbf{A}$ is constructed according to the lexicographic order of letters in the alphabet and has size $\kappa\ell \times \kappa\ell$, where the $\ell \times \ell$ block with row letter s and column letter t is

$$\mathbf{A}(s, t) = \begin{cases} \mathbf{R}, & \text{supp}(s) \cap \text{supp}(t) \neq \emptyset, \\ \mathbf{S}, & \text{supp}(s) \cap \text{supp}(t) = \emptyset. \end{cases} \quad (1)$$

For the degenerate-base alphabet $\Sigma_6^\dagger = \{A, A|C, C, G, G|T, T\}$, we have

$$\mathbf{A}^\dagger = \begin{bmatrix} \mathbf{R} & \mathbf{R} & \mathbf{S} & \mathbf{S} & \mathbf{S} & \mathbf{S} \\ \mathbf{R} & \mathbf{R} & \mathbf{R} & \mathbf{S} & \mathbf{S} & \mathbf{S} \\ \mathbf{S} & \mathbf{R} & \mathbf{R} & \mathbf{S} & \mathbf{S} & \mathbf{S} \\ \mathbf{S} & \mathbf{S} & \mathbf{S} & \mathbf{R} & \mathbf{R} & \mathbf{S} \\ \mathbf{S} & \mathbf{S} & \mathbf{S} & \mathbf{R} & \mathbf{R} & \mathbf{R} \\ \mathbf{S} & \mathbf{S} & \mathbf{S} & \mathbf{S} & \mathbf{R} & \mathbf{R} \end{bmatrix}.$$

This FSTD adjacency matrix is obtained without enumerating all length- ℓ forbidden substrings. To derive closed-form

formulas and identify optimal degenerate-base alphabets, we further compress $\mathbf{A}$ by a colour-refinement partition, as in Construction 2 below.

Construction 2 (Colour-refinement partition for construction of compressed matrix): Set $\pi = [C_1, \dots, C_m]$, an ordered partition of the degenerate-base alphabet Σ_κ into disjoint classes of degenerate-base letters. For a given ordered partition π , the run-extension profile of a letter s is $\mathbf{p}_\pi(s) = (p_1(s), \dots, p_m(s))$, where $p_b(s) = |\{t \in C_b : \text{supp}(s) \cap \text{supp}(t) \neq \emptyset\}|$. The colour-refinement partition is constructed in three phases:

- **Initialization:** The initial classes of π_0 consist of letters with equal $|\mathcal{N}(s)|$ from Definition 4.
- **Refinement:** Compute the run-extension profiles and split each class of degenerate-base letters into subclasses whose letters have identical profiles. When a class C_a is split, its subclasses replace C_a in the ordered partition, while the order of all other classes is retained.
- **Stabilization:** Repeat the refinement step until no further split occurs. Set $\pi^* = [C_1, \dots, C_\eta]$ for the coarsest stable colour-refinement partition, where $\eta = |\pi^*|$ is the number of classes of degenerate-base letters.

After stabilization, for $s \in C_a$, set $[\mathbf{P}]_{a,b} = p_b(s)$ and $[\mathbf{Q}]_{a,b} = |C_b|$. The rows and columns of $\mathbf{P}$ and $\mathbf{Q}$ follow the class order in π^* , so the construction of $\mathbf{A}_{\text{comp}}$ is uniquely determined.

Example 1 (Colour-refinement partition for $\Sigma_6^\dagger$): The refinement procedure is demonstrated using the six-letter degenerate-base alphabet

$$\Sigma_6^\dagger = \{A, A|C, C, G, G|T, T\}.$$

- 1) **Initialization.** The support-intersection set sizes from Definition 4 determine the initial classes. For this alphabet,

$$|\mathcal{N}(A)| = |\mathcal{N}(C)| = |\mathcal{N}(G)| = |\mathcal{N}(T)| = 2, \\ |\mathcal{N}(A|C)| = |\mathcal{N}(G|T)| = 3.$$

Therefore, we obtain the initial partition

$$\pi_0 = \{\{A, C, G, T\}, \{A|C, G|T\}\}.$$

- 2) **Refinement.** With respect to π_0 , the run-extension profiles are

$$\mathbf{p}_{\pi_0}(A) = \mathbf{p}_{\pi_0}(C) = \mathbf{p}_{\pi_0}(G) = \mathbf{p}_{\pi_0}(T) = (1, 1), \\ \mathbf{p}_{\pi_0}(A|C) = \mathbf{p}_{\pi_0}(G|T) = (2, 1).$$

Therefore, we obtain identical profiles within each class, and no class is split further.

- 3) **Stabilization.** No further splits occur, so the refinement stabilizes at $\pi^* = \pi_0$. The colour-refinement partition with $\eta = 4$ is given by

$$C_1 = \{A, C, G, T\}, \quad C_2 = \{A|C, G|T\}.$$

Therefore, we obtain

$$\mathbf{P}^\dagger = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}, \quad \mathbf{Q}^\dagger = \begin{bmatrix} 4 & 2 \\ 4 & 2 \end{bmatrix}.$$

Lemma 1 (Class-wise aggregation under the stabilized partition): Let $\pi^* = [C_1, \dots, C_\eta]$ be the ordered partition into classes of degenerate-base letters obtained from Construc-

tion 2. For $s \in C_a$ and $b \in \{1, \dots, \eta\}$, with $p_b(s)$ computed with respect to π^* , we have

$$\sum_{t \in C_b} \mathbf{A}(s, t) = p_b(s) \mathbf{R} + (|C_b| - p_b(s)) \mathbf{S}. \quad (2)$$

Moreover, for each given a and b , the value $p_b(s)$ is identical for all $s \in C_a$.

Proof: By the definition of $p_b(s)$ and (1), the sum over $t \in C_b$ contains $p_b(s)$ copies of $\mathbf{R}$ and $|C_b| - p_b(s)$ copies of $\mathbf{S}$, which proves (2). Since π^* is stabilized, all letters in the same class have identical run-extension profiles.

Theorem 1 (Construction of compressed matrix): For a degenerate-base alphabet Σ_κ with adjacency matrix $\mathbf{A}$ constructed by the auxiliary $\ell \times \ell$ block matrices in (1), we have the compressed matrix construction

$$\mathbf{A}_{\text{comp}} = \mathbf{P} \otimes (\mathbf{R} - \mathbf{S}) + \mathbf{Q} \otimes \mathbf{S} \in \mathbb{N}^{\eta \ell \times \eta \ell},$$

where $\mathbf{P}, \mathbf{Q} \in \mathbb{N}^{\eta \times \eta}$ are obtained from Construction 2. This compressed matrix satisfies $\rho(\mathbf{A}) = \rho(\mathbf{A}_{\text{comp}})$.

Proof: Set the class-expansion matrix $\mathbf{E} \in \{0, 1\}^{\kappa \ell \times \eta \ell}$ by

$$[\mathbf{E}]_{(s,r),(a,q)} = \begin{cases} 1, & s \in C_a \text{ and } r = q, \\ 0, & \text{otherwise.} \end{cases}$$

This matrix expands one compressed coordinate (a, q) to all auxiliary states (s, q) with $s \in C_a$. For $s \in C_a$, (2) and the construction of $\mathbf{P}$ and $\mathbf{Q}$ give

$$\sum_{t \in C_b} \mathbf{A}(s, t) = [\mathbf{P}]_{a,b} (\mathbf{R} - \mathbf{S}) + [\mathbf{Q}]_{a,b} \mathbf{S} = \mathbf{A}_{\text{comp}}(a, b). \quad (3)$$

Equivalently, (3) is the matrix identity

$$\mathbf{A} \mathbf{E} = \mathbf{E} \mathbf{A}_{\text{comp}}. \quad (4)$$

Indeed, for row (s, r) with $s \in C_a$ and column (b, q) , the entry of $\mathbf{A} \mathbf{E}$ equals the sum of the entries from (s, r) to all states (t, q) with $t \in C_b$. By identity (3), this is exactly the entry of $\mathbf{A}_{\text{comp}}$ from (a, r) to (b, q) , which is also the entry of $\mathbf{E} \mathbf{A}_{\text{comp}}$ at row (s, r) and column (b, q) . This proves matrix identity (4). Set $\mathbf{v}$ to be the positive Perron–Frobenius eigenvector of $\mathbf{A}_{\text{comp}}$ associated with $\rho(\mathbf{A}_{\text{comp}})$. Since every row of $\mathbf{E}$ contains exactly one non-zero entry, $\mathbf{E} \mathbf{v}$ is positive. From matrix identity (4), we have

$$\mathbf{A}(\mathbf{E} \mathbf{v}) = \mathbf{E} \mathbf{A}_{\text{comp}} \mathbf{v} = \rho(\mathbf{A}_{\text{comp}}) \mathbf{E} \mathbf{v}.$$

Therefore, we obtain that $\mathbf{A}$ has a positive eigenvector with eigenvalue $\rho(\mathbf{A}_{\text{comp}})$. By the Perron–Frobenius theorem for nonnegative adjacency matrices [16], [17], this eigenvalue is the spectral radius of $\mathbf{A}$. Therefore, we obtain $\rho(\mathbf{A}) = \rho(\mathbf{A}_{\text{comp}})$.

IV. CLOSED-FORM CONSTRAINED CAPACITY OF THE SIX-LETTER DEGENERATE-BASE ALPHABET

We first derive the closed-form constrained-capacity polynomial for

$$\Sigma_6^\dagger = \{\mathbf{A}, \mathbf{A}|\mathbf{C}, \mathbf{C}, \mathbf{G}, \mathbf{G}|\mathbf{T}, \mathbf{T}\}.$$

For $\Sigma_6^\dagger$, the coarsest stable colour-refinement partition is $C_1 = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$ and $C_2 = \{\mathbf{A}|\mathbf{C}, \mathbf{G}|\mathbf{T}\}$. Therefore, we obtain

$$\mathbf{P}^\dagger = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}, \quad \mathbf{Q}^\dagger = \begin{bmatrix} 4 & 2 \\ 4 & 2 \end{bmatrix}. \quad (5)$$

The characteristic polynomial is taken from the compressed matrix $\mathbf{A}_{\text{comp}}^\dagger$, whose degree is 2ℓ , rather than from the comprehensive FSTD adjacency matrix of size $6^\ell \times 6^\ell$.

The integer sequence F_j is specified by $F_0 = 0$, $F_1 = 1$, and

$$F_j = 2F_{j-1} + F_{j-2}. \quad (6)$$

In particular, (6) has $F_1 = 2F_0 + F_{-1}$, so $F_{-1} = 1$.

Proposition 1 (Reduction of $\chi_{\mathbf{A}_{\text{comp}}^\dagger}(x)$ to a 2×2 determinant): The characteristic polynomial of $\mathbf{A}_{\text{comp}}^\dagger$ satisfies

$$\chi_{\mathbf{A}_{\text{comp}}^\dagger}(x) = \det(x^\ell \mathbf{I}_2 - \sum_{j=1}^{\ell} x^{\ell-j} (\mathbf{P}^\dagger)^j (\mathbf{P}^\dagger)^{-1} (\mathbf{Q}^\dagger - \mathbf{P}^\dagger)). \quad (7)$$

Proof: We have

$$\mathbf{A}_{\text{comp}}^\dagger = \mathbf{P}^\dagger \otimes \mathbf{R} + (\mathbf{Q}^\dagger - \mathbf{P}^\dagger) \otimes \mathbf{S}.$$

The matrix $\mathbf{R}$ increases a run of length r to length $r + 1$ for $1 \leq r < \ell$, while $\mathbf{S}$ resets the run-length to 1. Set $\mathbf{v} = (\mathbf{v}_1, \dots, \mathbf{v}_\ell)^T$, where $\mathbf{v}_r \in \mathbb{R}^2$ is the two-entry vector associated with run-length r . Then $(x \mathbf{I}_{2\ell} - \mathbf{A}_{\text{comp}}^\dagger) \mathbf{v} = \mathbf{0}$ is equivalent to

$$\begin{cases} (x \mathbf{I}_2 - (\mathbf{Q}^\dagger - \mathbf{P}^\dagger)) \mathbf{v}_1 - \mathbf{P}^\dagger \mathbf{v}_2 = 0, \\ x \mathbf{I}_2 \mathbf{v}_r - \mathbf{P}^\dagger \mathbf{v}_{r+1} = 0, & 2 \leq r \leq \ell - 1, \\ x \mathbf{I}_2 \mathbf{v}_\ell = 0. \end{cases} \quad (8)$$

Therefore, we obtain the coefficient matrix

$$\begin{bmatrix} x \mathbf{I}_2 - (\mathbf{Q}^\dagger - \mathbf{P}^\dagger) & -\mathbf{P}^\dagger & 0 & \dots & 0 \\ 0 & x \mathbf{I}_2 & -\mathbf{P}^\dagger & \ddots & \vdots \\ 0 & 0 & x \mathbf{I}_2 & \ddots & 0 \\ \vdots & & & \ddots & \ddots \\ 0 & \dots & 0 & 0 & x \mathbf{I}_2 \end{bmatrix}. \quad (9)$$

Starting with the lower-right 2×2 block of (9) and applying the Schur-complement identity recursively to the preceding block rows, we obtain

$$\chi_{\mathbf{A}_{\text{comp}}^\dagger}(x) = x^{2(\ell-1)} \det(x \mathbf{I}_2 - \sum_{j=0}^{\ell-1} x^{-j} (\mathbf{P}^\dagger)^j (\mathbf{Q}^\dagger - \mathbf{P}^\dagger)). \quad (10)$$

Multiplying the matrix inside the determinant in (10) by $x^{\ell-1}$, we obtain

$$\begin{aligned} \chi_{\mathbf{A}_{\text{comp}}^\dagger}(x) &= \det(x^\ell \mathbf{I}_2 - \sum_{j=0}^{\ell-1} x^{\ell-1-j} \\ &\quad \times (\mathbf{P}^\dagger)^{j+1} (\mathbf{P}^\dagger)^{-1} (\mathbf{Q}^\dagger - \mathbf{P}^\dagger)). \end{aligned}$$

After replacing $j+1$ by r in the summation, we obtain identity (7).

Proposition 2 (Expansion for powers of $\mathbf{P}^\dagger$): For all $j \geq 1$, we have

$$(\mathbf{P}^\dagger)^j = F_j \mathbf{P}^\dagger + F_{j-1} \mathbf{I}_2. \quad (11)$$

Proof: From matrix identity (5), we obtain

$$(\mathbf{P}^\dagger)^2 = 2\mathbf{P}^\dagger + \mathbf{I}_2. \quad (12)$$

Equation (11) holds for $j = 1$. If identity (11) holds for j ,

then multiplying by $\mathbf{P}^\dagger$ and using identity (12) gives

$$(\mathbf{P}^\dagger)^{j+1} = (2F_j + F_{j-1})\mathbf{P}^\dagger + F_j\mathbf{I}_2.$$

By recurrence (6), this is $F_{j+1}\mathbf{P}^\dagger + F_j\mathbf{I}_2$, so identity (11) follows by induction.

Theorem 2 (Closed-form constrained-capacity polynomial for $\Sigma_6^\dagger$): The constrained capacity of $\Sigma_6^\dagger$ is $\log_2 \lambda_\ell$ by the constrained-system capacity formula $\log_2 \rho$ [11], where λ_ℓ is the largest real root of

$$\begin{aligned} \chi_{\mathbf{A}_{\text{comp}}^\dagger}(x) &= x^{2\ell} + \sum_{j=1}^{\ell} (F_{j-1} - 4F_j)x^{2\ell-j} \\ &\quad + \sum_{j=1}^{\ell} (-1)^{j+1} F_{\ell-j+1} x^{\ell-j}. \end{aligned} \quad (13)$$

Proof: From matrix identity (5), we have $(\mathbf{P}^\dagger)^{-1}(\mathbf{Q}^\dagger - \mathbf{P}^\dagger) = \begin{bmatrix} -1 & 0 \\ 4 & 1 \end{bmatrix}$. Proposition 2 implies

$$\begin{aligned} \sum_{j=1}^{\ell} x^{\ell-j} (\mathbf{P}^\dagger)^j &= \left(\sum_{j=1}^{\ell} F_j x^{\ell-j} \right) \mathbf{P}^\dagger \\ &\quad + \left(\sum_{j=1}^{\ell} F_{j-1} x^{\ell-j} \right) \mathbf{I}_2. \end{aligned} \quad (14)$$

Substituting identity (14) into Proposition 1, together with

$$\begin{aligned} \text{tr} \begin{bmatrix} -1 & 0 \\ 4 & 1 \end{bmatrix} &= 0, \quad \text{tr}(\mathbf{P}^\dagger \begin{bmatrix} -1 & 0 \\ 4 & 1 \end{bmatrix}) = 4, \\ \det \begin{bmatrix} -1 & 0 \\ 4 & 1 \end{bmatrix} &= -1, \quad \det(\mathbf{P}^\dagger) = -1, \end{aligned}$$

we obtain

$$\begin{aligned} \chi_{\mathbf{A}_{\text{comp}}^\dagger}(x) &= x^{2\ell} - 4x^\ell \sum_{j=1}^{\ell} F_j x^{\ell-j} \\ &\quad + \left(\sum_{j=1}^{\ell} F_j x^{\ell-j} \right)^2 - 2 \left(\sum_{j=1}^{\ell} F_j x^{\ell-j} \right) \\ &\quad \times \left(\sum_{j=1}^{\ell} F_{j-1} x^{\ell-j} \right) - \left(\sum_{j=1}^{\ell} F_{j-1} x^{\ell-j} \right)^2. \end{aligned} \quad (15)$$

The last three terms of (15) are

$$\begin{aligned} &\left(\sum_{j=1}^{\ell} F_j x^{\ell-j} \right)^2 - 2 \left(\sum_{j=1}^{\ell} F_j x^{\ell-j} \right) \left(\sum_{j=1}^{\ell} F_{j-1} x^{\ell-j} \right) \\ &\quad - \left(\sum_{j=1}^{\ell} F_{j-1} x^{\ell-j} \right)^2. \end{aligned}$$

From recurrence (6), we have $F_m - 2F_{m-1} = F_{m-2}$. For the coefficient of $x^{2\ell-j}$ in these three terms, we obtain

$$\begin{aligned} &\sum_{r=1}^{j-1} (F_r F_{j-r} - 2F_r F_{j-r-1} - F_{r-1} F_{j-r-1}) \\ &= \sum_{r=1}^{j-1} (F_r F_{j-r-2} - F_{r-1} F_{j-r-1}) = F_{j-1}. \end{aligned}$$

For the coefficient of $x^{\ell-j}$ in these three terms, we obtain

$$\begin{aligned} &\sum_{r=j}^{\ell} (F_r F_{\ell+j-r} - 2F_r F_{\ell+j-r-1} - F_{r-1} F_{\ell+j-r-1}) \\ &= F_\ell F_{j-2} - F_{j-1} F_{\ell-1} = (-1)^{j+1} F_{\ell-j+1}. \end{aligned} \quad (16)$$

The equality in (16) follows from $F_\ell F_{j-2} - F_{j-1} F_{\ell-1} = (-1)^{j+1} F_{\ell-j+1}$, obtained from $F_{r+1} = 2F_r + F_{r-1}$. The non-quadratic term in (15) contributes $-4F_j$ to the coefficient of $x^{2\ell-j}$ and contributes nothing to the coefficient of $x^{\ell-j}$. Therefore, we obtain coefficient $F_{j-1} - 4F_j$ for $x^{2\ell-j}$ and coefficient $(-1)^{j+1} F_{\ell-j+1}$ for $x^{\ell-j}$, which proves (13).

V. OPTIMAL DEGENERATE-BASE ALPHABET IDENTIFICATION

For six-letter degenerate-base alphabets, the dominance step removes alphabets whose larger support letters cannot increase constrained capacity. For example, $M_1 = \mathbf{C}|\mathbf{T}$ and $M_2 = \mathbf{C}|\mathbf{G}|\mathbf{T}$ are strictly dominated by $M_1 = \mathbf{C}|\mathbf{T}$ and $M_2 = \mathbf{C}|\mathbf{G}$, as the latter admits more valid ℓ -RLL sequences and higher constrained capacity. Therefore, we obtain the two non-dominated candidates after removing strictly dominated alphabets under the ℓ -RLL constraint:

$$\begin{aligned} \Sigma_6^\dagger &= \{\mathbf{A}, \mathbf{A}|\mathbf{C}, \mathbf{C}, \mathbf{G}, \mathbf{G}|\mathbf{T}, \mathbf{T}\}, \\ \Sigma_6^\ddagger &= \{\mathbf{A}, \mathbf{A}|\mathbf{C}, \mathbf{C}, \mathbf{C}|\mathbf{G}, \mathbf{G}, \mathbf{T}\}. \end{aligned}$$

For $\Sigma_6^\dagger$, the coarsest stable colour-refinement partition is $C_1 = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{T}\}$ and $C_2 = \{\mathbf{A}|\mathbf{C}, \mathbf{G}|\mathbf{T}\}$. Therefore, we obtain

$$\mathbf{P}^\dagger = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}, \quad \mathbf{Q}^\dagger = \begin{bmatrix} 4 & 2 \\ 4 & 2 \end{bmatrix}.$$

For $\Sigma_6^\ddagger$, the coarsest stable colour-refinement partition is

$$C_1 = \{\mathbf{T}\}, \quad C_2 = \{\mathbf{A}, \mathbf{G}\}, \quad C_3 = \{\mathbf{C}\}, \quad C_4 = \{\mathbf{A}|\mathbf{C}, \mathbf{C}|\mathbf{G}\}.$$

Therefore, we obtain

$$\mathbf{P}^\ddagger = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 1 & 1 & 2 \end{bmatrix}, \quad \mathbf{Q}^\ddagger = \begin{bmatrix} 1 & 2 & 1 & 2 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 1 & 2 \\ 1 & 2 & 1 & 2 \end{bmatrix}. \quad (17)$$

The two candidates are compared through $\rho(\mathbf{A}_{\text{comp}}^\dagger)$ and $\rho(\mathbf{A}_{\text{comp}}^\ddagger)$. Since the characteristic polynomials induce different recurrences, a direct termwise comparison is not available, so we use a trace method.

Proposition 3 (Trace expansion for the compressed matrix):

For $\mathbf{T} = \mathbf{P} \otimes (\mathbf{R} - \mathbf{S}) + \mathbf{Q} \otimes \mathbf{S}$ and every integer $k \geq 0$, we have

$$\text{tr}(\mathbf{T}^k) = \sum_{a=0}^k \binom{k}{a} \text{tr}(\mathbf{P}^a \mathbf{Q}^{k-a}) \text{tr}((\mathbf{R} - \mathbf{S})^a \mathbf{S}). \quad (18)$$

Proof: Expand $(\mathbf{P} \otimes (\mathbf{R} - \mathbf{S}) + \mathbf{Q} \otimes \mathbf{S})^k$ in tensor-product form. Since $\mathbf{S}$ is idempotent, $\mathbf{S}^{k-a} = \mathbf{S}$ for $k-a \geq 1$. Taking traces and using $\text{tr}(\mathbf{X} \otimes \mathbf{Y}) = \text{tr}(\mathbf{X}) \text{tr}(\mathbf{Y})$, we obtain formula (18).

Proposition 4 (Trace recurrences for the two six-letter

degenerate-base alphabets): For large k , we have

$$\begin{aligned}\text{tr}((\mathbf{A}_{\text{comp}}^\dagger)^k) &= 6^k - \binom{k}{\ell} 6^{k-\ell-1} c_\ell + o\left(\binom{k}{\ell} 6^{k-\ell-1}\right), \\ \text{tr}((\mathbf{A}_{\text{comp}}^\ddagger)^k) &= 6^k - \binom{k}{\ell} 6^{k-\ell-1} e_\ell + o\left(\binom{k}{\ell} 6^{k-\ell-1}\right).\end{aligned}$$

where $c_\ell = \mathbf{g}^T(\mathbf{P}^\dagger)^\ell \mathbf{1}_2$, $e_\ell = \mathbf{f}^T(\mathbf{P}^\ddagger)^\ell \mathbf{1}_4$, $\mathbf{Q}^\dagger = \mathbf{1}_2 \mathbf{g}^T$, $\mathbf{Q}^\ddagger = \mathbf{1}_4 \mathbf{f}^T$, $\mathbf{g} = [4, 2]^T$, and $\mathbf{f} = [1, 2, 1, 2]^T$. Moreover, $c_\ell = 2c_{\ell-1} + c_{\ell-2}$ with $c_0 = 6$ and $c_1 = 14$, while $e_\ell = 4e_{\ell-1} - 2e_{\ell-2} - e_{\ell-3}$ with $e_0 = 6$, $e_1 = 16$, and $e_2 = 50$.

Proof: Let $\mathbf{1}_n$ be all-one and let $\mathbf{e}_i$ be the i -th unit vector. Since $\mathbf{S} = \mathbf{1}_\ell \mathbf{e}_1^T$, we have

$$\text{tr}((\mathbf{R} - \mathbf{S})^a \mathbf{S}) = \mathbf{e}_1^T (\mathbf{R} - \mathbf{S})^a \mathbf{1}_\ell.$$

Moreover, $(\mathbf{R} - \mathbf{S})\mathbf{1}_\ell = -\mathbf{e}_\ell$ and $(\mathbf{R} - \mathbf{S})\mathbf{e}_r = \mathbf{e}_{r-1}$ for $2 \leq r \leq \ell$. Therefore, we obtain

$$\text{tr}((\mathbf{R} - \mathbf{S})^a \mathbf{S}) = \begin{cases} 1, & a = 0, \\ -1, & a = \ell, \\ 0, & 1 \leq a \leq \ell - 1. \end{cases} \quad (19)$$

Substituting identity (19) into formula (18), only the summands with $a = 0$ and $a = \ell$ are needed for the trace comparison. For either $\mathbf{Q} = \mathbf{Q}^\dagger$ or $\mathbf{Q} = \mathbf{Q}^\ddagger$, we have $\mathbf{Q} = \mathbf{1}_m \mathbf{w}^T$ with $\mathbf{w}^T \mathbf{1}_m = 6$, and therefore $\mathbf{Q}^r = 6^{r-1} \mathbf{Q}$ for every $r \geq 1$. The $a = 0$ summand in (18) is $\text{tr}(\mathbf{Q}^k) = 6^k$, while the $a = \ell$ summand is

$$-\binom{k}{\ell} \text{tr}(\mathbf{P}^\ell \mathbf{Q}^{k-\ell}) = -\binom{k}{\ell} 6^{k-\ell-1} \text{tr}(\mathbf{P}^\ell \mathbf{Q}).$$

For $\mathbf{Q}^\dagger = \mathbf{1}_2 \mathbf{g}^T$, we obtain

$$\text{tr}((\mathbf{P}^\dagger)^\ell \mathbf{Q}^\dagger) = \text{tr}((\mathbf{P}^\dagger)^\ell \mathbf{1}_2 \mathbf{g}^T) = \mathbf{g}^T (\mathbf{P}^\dagger)^\ell \mathbf{1}_2 = c_\ell.$$

The matrix $\mathbf{P}^\dagger$ satisfies $(\mathbf{P}^\dagger)^2 = 2\mathbf{P}^\dagger + \mathbf{I}_2$. Multiplying the left side and right side of (12) by $\mathbf{g}^T (\mathbf{P}^\dagger)^{\ell-2}$ and $\mathbf{1}_2$, we obtain $c_\ell = 2c_{\ell-1} + c_{\ell-2}$. The initial terms are $c_0 = \mathbf{g}^T \mathbf{1}_2 = 6$ and $c_1 = \mathbf{g}^T \mathbf{P}^\dagger \mathbf{1}_2 = 14$.

For $\mathbf{Q}^\ddagger = \mathbf{1}_4 \mathbf{f}^T$, the trace of $(\mathbf{P}^\ddagger)^\ell \mathbf{Q}^\ddagger$ is

$$\text{tr}((\mathbf{P}^\ddagger)^\ell \mathbf{Q}^\ddagger) = \mathbf{f}^T (\mathbf{P}^\ddagger)^\ell \mathbf{1}_4 = e_\ell.$$

Equation (17) decomposes $\mathbf{P}^\ddagger$ as

$$\mathbf{P}^\ddagger = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \Lambda \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 1 & 2 \end{bmatrix}. \quad (20)$$

Therefore, we obtain $(\mathbf{P}^\ddagger)^\ell = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \Lambda^\ell \end{bmatrix}$. Substituting this identity into $e_\ell = \mathbf{f}^T (\mathbf{P}^\ddagger)^\ell \mathbf{1}_4$, with $\mathbf{f} = [1, \mathbf{h}^T]^T$, $\mathbf{1}_4 = [1, \mathbf{u}^T]^T$, $\mathbf{h} = [2, 1, 2]^T$, and $\mathbf{u} = [1, 1, 1]^T$, gives $e_\ell = 1 + \mathbf{h}^T \Lambda^\ell \mathbf{u}$. Since $\Lambda^3 = 4\Lambda^2 - 2\Lambda - \mathbf{I}_3$, multiplying by $\mathbf{h}^T \Lambda^{\ell-3}$ and $\mathbf{u}$ gives $e_\ell = 4e_{\ell-1} - 2e_{\ell-2} - e_{\ell-3}$. We have $e_0 = 6$, $e_1 = 16$, and $e_2 = 50$.

Theorem 3 (Optimal six-letter degenerate-base alphabet identification using simple recurrence): For sufficiently large k , we have $\text{tr}((\mathbf{A}_{\text{comp}}^\dagger)^k) > \text{tr}((\mathbf{A}_{\text{comp}}^\ddagger)^k)$, and therefore $\rho(\mathbf{A}_{\text{comp}}^\dagger) > \rho(\mathbf{A}_{\text{comp}}^\ddagger)$. Consequently, $\Sigma_6^\dagger$ is the optimal six-letter degenerate-base alphabet.

Proof: Set $d_\ell = e_\ell - c_\ell$. Proposition 4 gives $d_1 = 2$, $d_2 = 16$, $d_3 = 80$, $d_4 = 334$, and

$$d_{\ell+1} = 4d_\ell - 2d_{\ell-1} - d_{\ell-2} + c_{\ell-1} + c_{\ell-2}.$$

Since $c_j > 0$, the induction hypothesis $d_j > 0$ and $d_j \geq 3d_{j-1}$ gives $d_{\ell+1} > 4d_\ell - 2d_{\ell-1} - d_{\ell-2} + c_{\ell-1} + c_{\ell-2} > 3d_\ell$. Hence $e_\ell > c_\ell$ for every $\ell \geq 1$, and Proposition 4 gives $\text{tr}((\mathbf{A}_{\text{comp}}^\dagger)^k) > \text{tr}((\mathbf{A}_{\text{comp}}^\ddagger)^k)$ for sufficiently large k .

VI. CONCLUSION

This work proposes a compressed matrix construction for the ℓ -RLL constraint over degenerate-base alphabets using run-extension and run-reset auxiliary matrices. The construction yields minimal-degree closed-form constrained-capacity formulations without exponential length- ℓ substring enumeration and an explicit $O(1)$ -time criterion for capacity-maximizing alphabet identification without solving capacity roots numerically. Future work will integrate these constrained codes with synchronization-error-correcting codes for DNA storage channels.

ACKNOWLEDGMENTS

This work was supported by the Research Program through the National Research Foundation of Korea under Grant NRF-2022M3C1A3081366 and NRF-2023R1A2C1003546.

REFERENCES

- [1] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.
- [2] N. Goldman *et al.*, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, 2013.
- [3] S. M. H. T. Yazdi *et al.*, "DNA-based storage: Trends and methods," *IEEE Trans. Mol., Biol., Multi-Scale Commun.*, vol. 1, no. 3, pp. 230–248, Sep. 2015.
- [4] Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, 2017.
- [5] Y. Choi *et al.*, "High information capacity DNA-based data storage with augmented encoding characters using degenerate bases," *Sci. Rep.*, vol. 9, no. 1, p. 6582, 2019.
- [6] L. Anavy, I. Vaknin, O. Atar, R. Amit, and Z. Yakhini, "Data storage in DNA with fewer synthesis cycles using composite DNA symbols," *Nature Biotechnol.*, vol. 37, no. 10, pp. 1229–1236, 2019.
- [7] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Sci. Rep.*, vol. 9, no. 1, p. 9663, 2019.
- [8] O. Milenkovic and C. Pan, "DNA-based data storage systems: A review of implementations and code constructions," *IEEE Trans. Commun.*, vol. 72, no. 7, pp. 3803–3828, Jul. 2024.
- [9] O. Sabary, H. M. Kiah, P. H. Siegel, and E. Yaakobi, "Survey for a decade of coding for DNA storage," *IEEE Trans. Mol., Biol., Multi-Scale Commun.*, vol. 10, no. 2, pp. 253–271, Jun. 2024.
- [10] T. T. Nguyen, C. Wang, K. Cai, K. A. S. Immink, Y. Zhang, and Z. Yakhini, "Constrained coding for composite DNA: Channel capacity and efficient constructions," in *Proc. IEEE ISIT*, 2025, pp. 1–6.
- [11] B. H. Marcus, R. M. Roth, and P. H. Siegel, "An introduction to coding for constrained systems," lecture notes, 2001.
- [12] Y. Shen, C. Shangguan, Z. Lin, and G. Ge, "Constrained coding bounds via Goulden–Jackson cluster theorem," *IEEE Trans. Inf. Theory*, vol. 71, no. 12, pp. 9359–9374, Dec. 2025.
- [13] P. H. Lundow, "Compression of transfer matrices," *Discrete Math.*, vol. 231, no. 1–3, pp. 321–329, 2001.
- [14] P. H. Lundow and K. Markström, "Exact and approximate compression of transfer matrices for graph homomorphisms," *LMS J. Comput. Math.*, vol. 11, pp. 1–14, 2008.
- [15] S. Kiefer and B. D. McKay, "The iteration number of colour refinement," in *Proc. ICALP*, 2020, pp. 73:1–73:19.
- [16] O. Perron, "Zur Theorie der Matrizen," *Math. Ann.*, vol. 64, no. 2, pp. 248–263, Jun. 1907.
- [17] G. Frobenius, "Über Matrizen aus nicht negativen Elementen," *Sitzungsber. Königl. Preuss. Akad. Wiss.*, pp. 456–477, 1912.