Reconstruction Codes for DNA Sequences with Uniform Tandem-Duplication Errors

Yonatan Yehezkeally, *Student Member, IEEE,* and Moshe Schwartz, *Senior Member, IEEE*

***Abstract*—DNA as a data storage medium has several advantages, including far greater data density compared to electronic media. We propose that schemes for data storage in the DNA of living organisms may benefit from studying the reconstruction problem, which is applicable whenever multiple reads of noisy data are available. This strategy is uniquely suited to the medium, which inherently replicates stored data in multiple distinct ways, caused by mutations. We consider noise introduced solely by uniform tandem-duplication, and utilize the relation to constant-weight integer codes in the Manhattan metric. By bounding the intersection of the cross-polytope with hyperplanes, we prove the existence of reconstruction codes with full rate, as well as suggest a construction for a family of reconstruction codes.**

***Index Terms*—DNA storage, reconstruction, string-duplication systems, tandem-duplication errors**

I. Introduction

DNA is attracting considerable attention in recent years as a medium for data storage, due to its high density and longevity [8]. Data storage in DNA may provide integral memory for synthetic-biology methods, where such is required, and offer a protected medium for long-period data storage [4], [42]. In particular, storage in the DNA of living organisms is now becoming feasible [40]; it has varied usages, including watermarking genetically modified organisms [3], [16], [35] or research material [21], [42], and even affords some concealment to sensitive information [9]. Naturally, therefore, data integrity in such media is of great interest.

Several recent works have studied the inherent constraints of storing and retrieving data from DNA. While desired sequences (over quaternary alphabet) may be synthesized (albeit, while suffering from substitution noise), generally data can only be read by observation of its subsequences, quite possibly an incomplete observation [22]. Moreover, the nature of DNA and current technology results in asymmetric errors which depend upon the dataset [14]. The medium itself also introduces other types of errors which are atypical in electronic storage, such as symbol/block-deletion and adjacent transpositions (possibly complemented) [15]. Finally, the purely combinatorial problem of recovering a sequence from the multiset of all its subsequences (including their numbers of incidence), was also studied, e.g., [1], [41], as well as coding schemes involving only these multisets (or their profile vectors – describing the incidence frequency of each subsequence) [38].

Other works were concerned with data storage in the DNA of a living organism. While this affords some level of protection to the data, and even propagation (through DNA replication), it is also exposed to specific noise mechanisms due to mutations. Examples of such noise include symbol insertions, deletion, substitutions (point-mutation), and duplication (including tandem- and interspersed-duplication). Therefore, schemes for data storage in live DNA must address data integrity and error-correction.

In an effort to better understand these typical noise mechanisms, their potential to generate the diversity observed in nature was studied. [12] classified the *capacity* and/or *expressiveness* of the systems of sequences over a finite alphabet generated by four distinct substring duplication rules: end-duplication, tandem-duplication, palindromic-duplication, and interspersed-duplication. [18] fully characterized the expressiveness of bounded tandem-duplication systems, proved bounds on their capacity (and, in some cases, even exact values). [20] later showed that when point-mutations act together with tandem-duplication as a sequence-generation process, they may actually increase the capacity of the generated system. [2] looked at the typical *duplication distance* of binary sequences; i.e., the number of tandem-duplications generating a binary sequence from its root. It was proven that for all but an exponentially small number of sequences that number is proportional to the sequence length. Further, when tandem-duplication is combined with point-mutations (here, only within the duplicated string), it was shown that the frequency of substitutions governs whether that distance becomes logarithmic.

The generative properties of interspersed-duplication were also studied from a probabilistic point of view. [11], [13] showed (under assumption of uniformity) that the frequencies of incidence for each subsequence converge to the same limit achieved by an i.i.d. source, thus reinforcing the notion that interspersed-duplication is–on its own–capable of generating diversity. [10] specifically looked at tandem- and end-duplication, and found exact capacities in the case of duplication length 1 by a generalization of the Pólya urn model that applies to strings. It also tightly bounded the capacity of complement tandem-duplication, a process where the duplicated symbol is complemented (using binary alphabet).

Finally, error-correcting codes for data affected by tandem-duplication have been studied in [19], which presented a

This work was presented in part at ISIT'2018.

This work was supported in part by the Israel Science Foundation (ISF) under grant no. 270/18.

The authors are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel (e-mail: yonatany@post.bgu.ac.il; schwartz@ee.bgu.ac.il).

construction of optimal-size codes for correcting any number of errors under *uniform tandem-duplication* (fixed duplication length), computing their (and thus, the optimal-) capacity. It also presented a framework for the construction of optimal codes for the correction of a fixed number of errors. Next, it studies bounded tandem-duplications, where a characterization of the capacity of error-correcting codes is made for small constants. In general, it characterized the cases where the process of tandem-duplication can be traced back uniquely. More recently, a flurry of activity in the subject includes works such as [27], [29], [30] which provide some implicit and explicit constructions for uniform tandem-duplication codes, as well as some bounds.

However, classical error-correction coding ignores some properties of the DNA storage channel; namely, stored information is expected to be replicated, even as it is mutated. This lends itself quite naturally to the reconstruction problem [34], which assumes that data is simultaneously transmitted over several noisy channels, and a decoder must therefore estimate that data based on several (distinct) noisy versions of it. Solutions to this problem have been studied in several contexts. It was solved in [34] for sequence reconstruction over finite alphabets, where several error models were considered, such as substitutions, transpositions and deletions. Moreover, a framework was presented for solving the reconstruction problem in general cases of interest in coding theory, utilizing a graph representation of the error model, which was further developed in [32], [33]. The problem was also studied in the context of permutation codes with transposition and reversal errors [23]–[25], and partially solved therein. Later, applications were found in storage technologies [6], [7], [43], [44], since modern application might preclude the retrieval of a single data point, in favor of multiple-point requests. However, the problem hasn't been addressed yet for data storage in the DNA of living organisms, where it may be most applicable.

In this paper, we study the reconstruction problem over DNA sequences, with uniform tandem-duplication errors. The main contributions of the paper are the following: We show that reconstruction codes in this setting are necessarily error-correcting codes with appropriately chosen minimum distance, based on the uncertainty parameter. We also show that in two asymptotic regimes, we can always obtain higher size than error-correcting codes. These asymptotic regimes include what we believe is the most interesting one, where the uncertainty is sublinear, and the time (number of mutations) is bounded by a constant.

The paper is organized as follows: In Section II we present notations and definitions. In Section III we demonstrate that reconstruction codes partition into error-correcting codes and find the requisite minimal-distance of each part, as a function of the reconstruction parameters. We see that these parts can be isometrically embedded as constant-weight codes in the Manhattan metric. Finally, in Section IV we show that reconstruction codes exist with full capacity, and also suggest a construction for reconstruction codes; we also briefly review recent results, published after the submission of this paper. We conclude with closing remarks in Section V.

II. Preliminaries

Throughout this paper, though DNA is composed of four nucleotide bases, we observe the more general case of sequences over a finite alphabet; since the alphabet elements are immaterial to our discussion, we denote it throughout as $\mathbb{Z}_q$. We observe the set of finite sequences (also: *words*) over it $\mathbb{Z}_q^* \triangleq \bigcup_{n=0}^{\infty} \mathbb{Z}_q^n$. For any two words $u, v \in \mathbb{Z}_q^*$, we denote their concatenation uv. For each word $x \in \mathbb{Z}_q^n$, we denote its *length* $|x| = n$. We also take special note of the set of words with length higher than or equal to some $0 < k \in \mathbb{N}$, which we denote $\mathbb{Z}_q^{\geqslant k} \triangleq \left\{ x \in \mathbb{Z}_q^* \;\middle|\; |x| \geqslant k \right\}$. For ease of notation, we let $\mathbb{N}$ stand for the set of non-negative integers throughout the paper; when an integer is assumed to be strictly positive, we make special note of that fact.

For $0 < k \in \mathbb{N}$, $i \in \mathbb{N}$, we define a *tandem-duplication* of *duplication-length* k by the mappings

$$\mathcal{T}_{k,i}(x) \triangleq \begin{cases} uvvw & x = uvw,\ |u| = i, |v| = k, \\ x & \text{otherwise.} \end{cases}$$

If $y = \mathcal{T}_{k,i}(x)$ and $y \neq x$ (which occurs whenever $|x| \geqslant i + k$), we say that y is a *descendant* of x, and denote $x \underset{k}{\Longrightarrow} y$. In what follows, we focus on the uniform tandem-duplication model (i.e., we fix k) because of its simplicity.

Further, given a sequence $\{x_j\}_{j=0}^{t} \subseteq \mathbb{Z}_q^*$ such that for all $0 \leqslant j < t$, $x_j \underset{k}{\Longrightarrow} x_{j+1}$, we say that x_t is a *t-descendant* of x_0, and denote $x_0 \overset{t}{\underset{k}{\Longrightarrow}} x_t$. For completeness, we also denote $x \overset{0}{\underset{k}{\Longrightarrow}} x$. Finally, if there exists some $t \in \mathbb{N}$ such that $x \overset{t}{\underset{k}{\Longrightarrow}} y$, we also denote $x \overset{*}{\underset{k}{\Longrightarrow}} y$.

We denote the set of t-descendants of $x \in \mathbb{Z}_q^*$ as

$$D_k^t(x) \triangleq \left\{ y \in \mathbb{Z}_q^* \;\middle|\; x \overset{t}{\underset{k}{\Longrightarrow}} y \right\},$$

for some $t \in \mathbb{N}$. We also denote the *descendant cone* of x by $D_k^*(x) \triangleq \bigcup_{t=0}^{\infty} D_k^t(x)$.

We say that $x \in \mathbb{Z}_q^{\geqslant k}$ is *irreducible* if $x \in D_k^*(y)$ implies $y = x$. We exclude from the definition shorter words, for which the condition vacuously holds. We denote by Irr_k the set of all irreducible words, and $\mathrm{Irr}_k(n) \triangleq \mathrm{Irr}_k \cap \mathbb{Z}_q^n$.

It was shown in [20], [31] that for each word $x \in \mathbb{Z}_q^{\geqslant k}$, a unique irreducible word exists for which x is a descendant. We call it the *root* of x, and denote it by $R_k(x)$. This induces an equivalence relation by $x \sim_k y$ if $R_k(x) = R_k(y)$.

We also follow [20] in defining, for $x \in \mathbb{Z}_q^{\geqslant k}$, $\mathrm{Pref}_k(x)$ as the length-k *prefix* of x, and $\mathrm{Suff}_k(x)$ as its suffix; i.e., if $x = uu' = v'v$ where $|u| = |v| = k$, then $\mathrm{Pref}_k(x) = u$ and $\mathrm{Suff}_k(x) = v$. Using this notation, we define an embedding $\phi_k : \mathbb{Z}_q^{\geqslant k} \to \mathbb{Z}_q^k \times \mathbb{Z}_q^*$ by

$$\phi_k(x) \triangleq \left(\mathrm{Pref}_k(x), \mathrm{Suff}_{|x|-k}(x) - \mathrm{Pref}_{|x|-k}(x) \right).$$

It is seen in [20] that this mapping is indeed injective. Further, it was shown that, defining $\zeta_{k,i} : \mathbb{Z}_q^k \times \mathbb{Z}_q^* \to \mathbb{Z}_q^k \times \mathbb{Z}_q^*$ by

$$\zeta_{k,i}(a, b) \triangleq \begin{cases} (a, b_1 0^k b_2) & b = b_1 b_2,\ |b_1| = i, \\ (a, b) & \text{otherwise,} \end{cases}$$

where $0 < k \in \mathbb{N}$, $i \in \mathbb{N}$, we have

$$\phi_k(\mathcal{T}_{k,i}(x)) = \zeta_{k,i}(\phi_k(x)).$$

The simplicity of $\zeta_{k,i}$ in comparison to $\mathcal{T}_{k,i}$ motivates the analysis of tandem-duplications using the ϕ_k images of sequences.

If $b \in \mathbb{Z}_q^*$ is composed of the subsequences

$$b = 0^{s_1} w_1 0^{s_2} \cdots w_m 0^{s_{m+1}}; \quad w_1, \dots, w_m \in \mathbb{Z}_q \setminus \{0\}$$

we define

$$\mu(b) \triangleq 0^{s_1 \bmod k} w_1 0^{s_2 \bmod k} \cdots w_m 0^{s_{m+1} \bmod k},$$
$$\sigma(b) \triangleq \left(\left\lfloor \frac{s_1}{k} \right\rfloor, \dots, \left\lfloor \frac{s_{m+1}}{k} \right\rfloor \right).$$

We may note that $\mathrm{wt}_H(b) = \mathrm{wt}_H(\mu(b)) = m$, where wt_H is the Hamming weight, and $\sigma(b) \in \mathbb{N}^{\mathrm{wt}_H(b)+1} = \mathbb{N}^{\mathrm{wt}_H(\mu(b))+1}$. We also observe that b is recoverable from $\sigma(b), \mu(b)$. It was proven in [20, Cor. 10] that if $\phi_k(x) = (a, b)$ then

$$\phi_k(R_k(x)) = (a, \mu(b)).$$

Thus, if $x, y \in \mathbb{Z}_q^{\geqslant k}$, $\phi_k(x) = (a_1, b_1)$ and $\phi_k(y) = (a_2, b_2)$, then $x \sim_k y$ if and only if $a_1 = a_2$ and $\mu(b_1) = \mu(b_2)$. Moreover, $x \in \mathrm{Irr}_k$ if and only if $\sigma(b_1) = (0, 0, \dots, 0)$. Note that, equivalently, we may say that b contains no zero-runs of length k; such sequences are called $(0, k-1)_q$-*Run-Length-Limited*, or $(0, k-1)_q$-RLL.

For $x \in \mathrm{Irr}_k$, $\phi_k(x) = (a, b)$, we denote $m(x) \triangleq \mathrm{wt}_H(b)$ and define $\psi_x : D_k^*(x) \to \mathbb{N}^{m(x)+1}$ by $\psi_x(y) \triangleq \sigma(b')$, where $\phi_k(y) = (a, b')$.

Finally, for $n \geqslant k$ and $x, y \in \mathbb{Z}_q^n$ we define

$$d_k(x, y) \triangleq \min\left\{ t \in \mathbb{N} \;\middle|\; D_k^t(x) \cap D_k^t(y) \neq \emptyset \right\},$$

or $d_k(x, y) = \infty$ if $\{t \in \mathbb{N} \mid D_k^t(x) \cap D_k^t(y) \neq \emptyset\} = \emptyset$. It was shown in [20, Lem. 14] that $d_k(x, y) = \infty$ if and only if $x \not\sim_k y$, hence $d_k(\cdot, \cdot)$ is finite on $D_k^t(x)$, for any particular $x \in \mathbb{Z}_q^{\geqslant k}$. Furthermore, [20, Lem. 19] shows that for any $x \sim_k y$ with $|x| = |y|$ it holds that

$$d_k(x, y) = \tfrac{1}{2} \left\| \sigma(b_1) - \sigma(b_2) \right\|_1,$$

thus $d_k(\cdot, \cdot)$ defines a metric on each equivalence class of $\sim_k$.

III. Reconstruction Codes

The reconstruction problem in the context of uniform tandem-duplication errors can be stated as follows: suppose data is encoded in $C \subseteq \mathbb{Z}_q^n$, and suppose we later are able to read distinct $x_0, x_1, \dots, x_N \in D_k^t(c)$ for some specific $c \in C$ and $t \in \mathbb{N}$; can we uniquely identify c?

It is apparent (see [34]) that to allow successful reconstruction we require codes to satisfy the following.

Definition 1 Take $N, t, n > 0$. We say that $C \subseteq \mathbb{Z}_q^n$ is a *uniform tandem-duplication reconstruction code*, which we abbreviate as an $(N, t, k)_q$-UTR code, if

$$\max\left\{ |D_k^t(c) \cap D_k^t(c')| \;\middle|\; c, c' \in C,\ c \neq c' \right\} \leqslant N.$$

The purpose of this section is to characterize reconstruction codes. By an evaluation of the size of intersection of descendant cones, we determine the achievable size of $(N, t, k)_q$-UTR codes. We shall state the solution to this problem in terms of error-correcting codes for the Manhattan metric, and devote the next section to an observation of such codes.

A. Structure of descendant cones

Throughout this section we fix some $x \in \mathrm{Irr}_k$, and denote $\phi_k(x) = (a, b)$.

As noted above, for all $y \in D_k^*(x)$, we have $\phi_k(y) = (a, b')$, with $\mu(b') = b$ (hence, in particular, $\mathrm{wt}_H(b') = \mathrm{wt}_H(b)$). We therefore denote $m = m(x) = \mathrm{wt}_H(b)$ and make the following definition:

Definition 2 We let $\psi_x : D_k^*(x) \to \mathbb{N}^{m+1}$ be defined by $\psi_x(y) = \sigma(b')$, where $\phi_k(y) = (a, b')$.

It was noted in the previous section that ψ_x is then distance-preserving from $(D_k^*(x), d_k)$ to $\left(\mathbb{N}^{m+1}, \frac{1}{2} \|\cdot\|_1\right)$ (the definition of d_k, made here specifically for sequences of equal-length, can be extended to $D_k^*(x)$ by considering the shortest path between any two sequences, but for simplicity in what follows, we shall implicitly only consider d_k as a metric over $D_k^t(x)$ for any given $t \in \mathbb{N}$).

Definition 3 We define on $\mathbb{N}^{m+1}$ the partial order $b' \preceq b''$ if for every coordinate $i = 1, \dots, m+1$ it holds that $b'_i \leqslant b''_i$ ($\preceq$ is the well-known product order).

The poset $\left(\mathbb{N}^{m+1}, \preceq\right)$ has a simple structure. We shall therefore find it more convenient to consider $D_k^*(x)$ in these terms:

Lemma 4 ψ_x is a poset isomorphism from $\left(D_k^*(x), \xRightarrow[k]{*}\right)$ to $\left(\mathbb{N}^{m+1}, \preceq\right)$. In particular,

1) For all $y, y' \in D_k^*(x)$ there exists $z \in D_k^*(y) \cap D_k^*(y')$ such that
$$D_k^*(y) \cap D_k^*(y') = D_k^*(z);$$

2) If in addition $|y| = |y'|$ then for all $t \in \mathbb{N}$
$$\left| D_k^t(y) \cap D_k^t(y') \right| = \begin{cases} 0 & t < d_k(y, y'), \\ \left| D_k^{t - d_k(y,y')}(x) \right| & t \geqslant d_k(y, y'). \end{cases}$$

Proof: We note that $x \in \mathrm{Irr}_k$, hence $\psi_x(x) = (0, 0, \dots, 0) \in \mathbb{N}^{m+1}$. Further, we note that in the image of ϕ_k, a tandem-duplication $\zeta_{k,i}$ corresponds to increasing by one a single coordinate of $\sigma(\cdot)$, i.e., an addition of a unit vector $e_j \in \mathbb{N}^{m+1}$ to $\psi_x(\cdot)$.

Hence, ψ_x is indeed a poset isomorphism, and we see that $\xRightarrow[k]{*}$ endows $D_k^*(x)$ with a lattice structure; We denote the *join* of $y, y' \in D_k^*(x)$ as $y \vee y'$, and their *meet* $y \wedge y'$. It follows that $z = y \vee y'$ satisfies item 1.

Finally, if $|y| = |y'|$ then by definition of $d \triangleq d_k(y, y')$ we have $z = y \vee y' \in D_k^d(y) \cap D_k^d(y')$, and item 2 is now straightforward to prove from the poset-isomorphism. ■

Given Lemma 4, we can now find the size of intersection of descendant cones for any $c, c' \in \mathbb{Z}_q^n$ ($n \geqslant k$), keeping in mind that $D_k^*(c) \cap D_k^*(c') \neq \emptyset$ if and only if $c \sim_k c'$.

Lemma 5 For $x \in \mathrm{Irr}_k$, $|D_k^t(x)| = \binom{t + m(x)}{m(x)}$.

Proof: By Lemma 4 we know that

$$D_k^t(x) = \{ y \in D_k^*(x) \mid \|\psi_x(y)\|_1 = t \}.$$

Since $\psi_x : D_k^*(x) \to \mathbb{N}^{m(x)+1}$ is bijective, $|D_k^t(x)|$ equals the number of distinct integer solutions to $\sum_{j=1}^{m+1} x_j = t$, where $x_1, \dots, x_{m+1} \geqslant 0$ (equivalently, the number of distinct ways to distribute t identical balls into $m(x)+1$ bins). ■

B. Size of reconstruction codes

In this section we aim to estimate the maximal size of $(N,t,k)_q$-UTR codes.

Definition 6 For $m, r > 0$ we denote the *simplex* of *dimension* m and *weight* r, or $(m,r)-$simplex

$$\Delta_r^m \triangleq \left\{ (x_i)_{i=1}^{m+1} \in \mathbb{N}^{m+1} \;\middle|\; \sum_{j=1}^{m+1} x_j = r \right\}.$$

Theorem 7 We take positive integers N, t and $n > k$. For $C \subseteq \mathbb{Z}_q^n$ and $x \in \mathrm{Irr}_k$ we partition $C_x \triangleq C \cap D_k^*(x)$ and define $r(x) \triangleq \frac{n-|x|}{k}$.

If $C_x \neq \emptyset$ then $r(x) \in \mathbb{N}$ and $r(x) < \left\lfloor \frac{n}{k} \right\rfloor$. Moreover, C is an $(N,t,k)_q$-UTR code if and only if for all $x \in \mathrm{Irr}_k$ such that $C_x \neq \emptyset$, the image $\psi_x(C_x) \subseteq \Delta_{r(x)}^{m(x)}$ satisfies

$$\min\left\{ \tfrac{1}{2} \left\| c - c' \right\|_1 \;\middle|\; c \neq c' \in \psi_x(C_x) \right\} \geqslant d_{N,t}(m(x)),$$

where we make the notation

$$d_{N,t}(m) \triangleq \min\left\{ \delta \in \mathbb{N} \;\middle|\; \binom{t-\delta+m}{m} \leqslant N \right\}.$$

Proof: If $C \cap D_k^*(x) \neq \emptyset$ then it follows from the definitions that for some $r \in \mathbb{N}$ we have $|x| + rk = n$; since $|x| \geqslant k$, necessarily $r = r(x) < \left\lfloor \frac{n}{k} \right\rfloor$. Furthermore, $C \cap D_k^*(x) = C \cap D_k^r(x)$, hence we have seen in the proof of Lemma 4 that for all $y \in D_k^r(x)$ we have $\psi_x(y) = \sum_{u=1}^{r} e_{j_u} \in \Delta_r^{m(x)}$.

In addition, by Lemma 4 and Lemma 5, for all $x \in \mathrm{Irr}_k$ and $y \neq y' \in C_x$ the size of intersection $D_k^t(y) \cap D_k^t(y')$ is $\binom{t-d_k(y,y')+m(x)}{m(x)}$. It follows that C_x is an $(N,t,k)_q$-UTR code if and only if that size is no greater than N for all such $y, y' \in C_x$.

Recalling that ψ_x is bijective and distance-preserving, i.e., that $d_k(y,y') = \frac{1}{2} \left\| \psi_x(y) - \psi_x(y') \right\|_1$, the claim follows for C_x.

To conclude the proof, we recall that for $x, x' \in \mathrm{Irr}_k$ we have $D_k^*(x) \cap D_k^*(x') = \emptyset$, hence C is an $(N,t,k)_q$-UTR if and only if the same is true for C_x, for all $x \in \mathrm{Irr}_k$. ■

In other words, Theorem 7 states that the intersection of a uniform-tandem-duplication reconstruction code C with the descendant cone of any irreducible word $D_k^*(x)$ can be viewed as an error-correcting code with a suitable minimal distance. Further, we see that these error-correcting codes are equivalent to codes in the Manhattan metric over a simplex $\Delta_{r(x)}^{m(x)}$. We note here, however, that this does not hold for C in general: not only is each code's minimal distance dependent on x, but the dimension and weight of the simplex in which that code exists do, as well.

We therefore see that constructions and bounds on the size of error-correcting codes for uniform tandem-duplication depend on doing the same for error-correcting codes in the Manhattan metric over Δ_r^m. We start by notating the maximal size of such codes:

Definition 8 For $m, r > 0$ and $d \geqslant 0$ we define

$$M(m,r,d) \triangleq \max\left\{ |C| \;\middle|\; C \subseteq \Delta_r^m,\, \min_{\substack{c,c' \in C \\ c \neq c'}} \tfrac{1}{2} \left\| c - c' \right\|_1 \geqslant d \right\}.$$

We now reiterate that if $C \subseteq \mathbb{Z}_q^n$, $x, x' \in \mathrm{Irr}_k(n-rk)$ (i.e., $r(x) = r(x') = r$) and $m(x) = m(x')$, then $D_k^{n-rk}(x) \cong D_k^{n-rk}(x')$ (through, e.g., $\psi_{x'}^{-1} \circ \psi_x$). It is therefore practical to assume $|C_x| = |C_{x'}| = M(m,r,d_{N,t}(m))$ for all such x, x'. This results in the following corollary, which concludes this section:

Corollary 9 If $C \subseteq \mathbb{Z}_q^n$ is an $(N,t,k)_q$-UTR code, and for all $x \in \mathrm{Irr}_k$ it holds that $|C_x| = M(m,r,d_{N,t}(m))$, then

$$\begin{aligned} |C| = \sum_{r=0}^{\lfloor n/k \rfloor - 1} \sum_{m} & M(m,r,d_{N,t}(m)) \cdot \\ & \cdot \left| \left\{ x \in \mathrm{Irr}_k(n-rk) \mid m(x) = m \right\} \right| \\ = \sum_{r=0}^{\lfloor n/k \rfloor - 1} \sum_{m} & M(m,r,d_{N,t}(m)) \cdot q^k \cdot \\ & \cdot \left| \left\{ b \in \mathbb{Z}_q^{n-(r+1)k} \;\middle|\; \substack{b \text{ is } (0,k-1)_q\text{-RLL} \\ \mathrm{wt}_H(b)=m} \right\} \right| \end{aligned}$$

Proof: First, trivially, $|C| = \sum_{x \in \mathrm{Irr}_k} |C_x|$.

Observe that $x \in \mathrm{Irr}_k$ satisfies $C_x \neq \emptyset$, $r(x) = r$ and $m(x) = m$, if and only if $x \in \mathbb{Z}_q^{n-rk}$ and in $\phi_k(x) = (a,b)$, b is $(0,k-1)_q$-RLL, and $\mathrm{wt}_H(b) = m$.

The rest now follows from Theorem 7. ■

Corollary 9 motivates us to estimate the optimal size of error-correcting codes in the Manhattan metric over the (m,r)-simplex. This topic was examined in some depth in [28], where a construction based on Sidon sets (of particular interest for our application, see [26], and references therein) was proposed, leading to lower bounds tighter than the Gilbert-Varshamov bound. For our purposes, we cite an asymptotic result (we slightly rephrase):

Lemma 10 [28, Eq. 36] Take $\mu \in (0,1)$, $\rho > 0$ and integer sequences $(m_n)_{n>0}$, $(r_n)_{n>0}$ such that $\lim_{n\to\infty} \frac{m_n}{n} = \mu$ and $\lim_{n\to\infty} \frac{r_n}{n} = \rho$. Also take a fixed $d > 0$. Then

$$\lim_{n\to\infty} \frac{1}{n} \log_2 M(m_n, r_n, d) = (\mu + \rho) H\left(\frac{1}{1 + \frac{\rho}{\mu}} \right). \quad (1)$$

C. Minimal distance of reconstruction codes

Next, before we can ascertain the sizes of error-correcting codes over simplices, we bound their requisite minimal distance. That is, given $N, t > 0$ and $m > 0$, we establish bounds on

$$d_{N,t}(m) \triangleq \min\left\{ \delta \in \mathbb{N} \;\middle|\; \binom{t-\delta+m}{m} \leqslant N \right\}$$

seen in Theorem 7.

Lemma 11 If $N \leqslant m$ then $d_{N,t}(m) = t$.

Proof: We may verify by substitution that $\delta = t$ satisfies $\binom{t-\delta+m}{m} \leqslant N$, while $\delta = t-1$ does not. Using the strict monotonicity of $s \mapsto \binom{s+m}{m}$, we are done. ■

In order to find a practical bound for $d_{N,t}(m)$ when $N > m$, we first require the following three lemmas:

Lemma 12 1) [36, Ch.10, Sec.11, Lem.7] For integers $0 < k < n$ it holds that

$$\sqrt{\frac{n}{8k(n-k)}}2^{nH\left(\frac{k}{n}\right)} \leqslant \binom{n}{k} \leqslant \sqrt{\frac{n}{2\pi k(n-k)}}2^{nH\left(\frac{k}{n}\right)}$$

where H is the binary entropy function, defined by $H(p) \triangleq -p\log_2 p - (1-p)\log_2(1-p)$.

2)

$$nH\left(\frac{k}{n}\right) - \frac{1}{2}\log_2(2n) \leqslant \log_2\binom{n}{k} < nH\left(\frac{k}{n}\right).$$

Proof: For item 2, we see that if $0 < k < n$ we have $n-1 \leqslant k(n-k) \leqslant \frac{n^2}{4}$, hence

$$\frac{n}{2\pi k(n-k)} \leqslant \frac{1}{2\pi}\left(1 + \frac{1}{n-1}\right) \leqslant \frac{1}{\pi} < 1,$$

$$\frac{n}{8k(n-k)} \geqslant \frac{1}{2n}.$$

Thus the claim trivially follows from item 1. ■

For ease of notation in what follows, we make the notation, for $1 \leqslant x \in \mathbb{R}$:

$$\mathcal{H}(x) \triangleq xH\left(\frac{1}{x}\right).$$

Lemma 13 For $N > m > 0$ and $t > 0$ it holds that

$$d_{N,t}(m) \leqslant \min\left\{\delta \in \mathbb{N} \;\middle|\; \mathcal{H}\left(1 + \frac{t-\delta}{m}\right) \leqslant \frac{\log_2 N}{m}\right\}.$$

Proof: Under the assumption, $\delta = t-1$ satisfies the inequality $\binom{t-\delta+m}{m} \leqslant N$. Therefore we may restrict the minimum to $\delta < t$, giving $0 < m < (t-\delta)+m$. Now, Lemma 12 implies

$$\log_2\binom{t-\delta+m}{m} \leqslant m\left(1 + \frac{t-\delta}{m}\right)H\left(\frac{1}{1+\frac{t-\delta}{m}}\right),$$

which completes the proof. ■

Lemma 14 For $x \geqslant 1$ it holds that $\mathcal{H}(x) \leqslant 2\sqrt{x-1}$.

Proof: The claim can be restated by the substitution $p = \frac{1}{x}$ as the known inequality $H(p)^2 \leqslant 4p(1-p)$ (its proof follows elementary calculus, and is omitted here). ■

Finally,

Theorem 15 Take $N > m > 0$. Then

$$d_{N,t}(m) \leqslant \max\left\{1, t - \left\lfloor\frac{(\log_2 N)^2}{4m}\right\rfloor\right\}.$$

Proof: Using Lemma 14 we may bound $\mathcal{H}\big(1 + \frac{t-\delta}{m}\big) \leqslant 2\sqrt{\frac{t-\delta}{m}}$. Lemma 13 therefore implies that it suffices to require $2\sqrt{\frac{t-\delta}{m}} \leqslant \frac{\log_2 N}{m}$, and reordering the inequality we get $\delta \geqslant t - \frac{(\log_2 N)^2}{4m}$, yielding the claim. ■

IV. Capacity of reconstruction codes

Definition 16 We define the *rate* of a code $C \subseteq \mathbb{Z}_q^n$ as

$$R(C) \triangleq \frac{1}{n}\log_q|C|,$$

and the capacity of a system $\mathcal{C} \subseteq \mathbb{Z}_q^*$ as

$$\operatorname{cap}(\mathcal{C}) \triangleq \limsup_{n\to\infty}\frac{1}{n}\log_q\left|\mathcal{C}\cap\mathbb{Z}_q^n\right|.$$

We are interested in $\sup\{\operatorname{cap}(\mathcal{C})\}$, where $\mathcal{C}$ is any family of reconstruction codes (i.e., $\mathcal{C}\cap\mathbb{Z}_q^n$ is an $(N_n, t_n, k)_q$-code for all n).

The purpose of this section is to determine that optimal capacity in two asymptotic regimes:

Regime I When $N_n = o(n)$ and $t_n = t$ is fixed.

Regime II When $N_n = 2^{\alpha n}$ and $t_n = \beta n$ for constants $\alpha, \beta > 0$ (such that $N_n, t_n \in \mathbb{N}$ for some, hence infinitely many, indices).

In practical applications, Regime I is likely to apply, since we may indeed expect the number of duplications t, which is dependent on the period of time before data is read, to be fixed w.r.t. n. The allowed uncertainty N_n will also likely be bounded. Regime II requires Theorem 15 (and some restrictions over the values of α, β), but allows us to calculate capacity in much the same way, which we do after presenting the first.

Note, since [19] showed that $\operatorname{Irr}_k(n)$ can correct any number of tandem-duplication errors, they are trivially $(N, t, k)_q$-codes for all N, t (more precisely, they are $(0, t, k)_q$-codes for all t). In comparison, in the setting we consider only t tandem-duplications are assumed to have occurred, therefore the codes we seek are less restrictive. Nevertheless, at the time of this paper's submission no bounds on the size of error-correcting codes for a fixed number of tandem-duplications were known; It is our purpose, then, to demonstrate that reconstruction codes exist which have strictly higher capacity than Irr_k, and suggest constructions for families of such codes.

First, we denote for any $n, r \in \mathbb{N}$ such that $n \geqslant k$ and $r < \left\lfloor\frac{n}{k}\right\rfloor$, and any $N, t \in \mathbb{N}$

$$\mathcal{M}_{N,t}(n,r) \triangleq \sum_m M(m, r, d_{N,t}(m))\cdot \\ \cdot\left|\left\{b \in \mathbb{Z}_q^{n-(r+1)k} \;\middle|\; \substack{b \text{ is } (0,k-1)_q\text{-RLL} \\ \mathrm{wt}_H(b)=m}\right\}\right|.$$

We recall for all n, if $r_n = \arg\max_r \mathcal{M}_{N,t}(n,r)$, that by Corollary 9 we have an $(N, t, k)_q$-code $C \subseteq \mathbb{Z}_q^n$ with $|C| \geqslant q^k\mathcal{M}_{N,t}(n, r_n)$. Corollary 9 also implies that for all $C \subseteq \mathbb{Z}_q^n$ it holds that $|C| \leqslant \frac{n}{k}q^k\mathcal{M}_{N,t}(n, r_n)$. We therefore focus on maximizing $\limsup_{n\to\infty}\frac{1}{n}\log_q\mathcal{M}_{N,t}(n, r_n)$ by choice of r_n.

In what follows, we take $\gamma \in (0,1)$ and set $r_n = \frac{1-\gamma}{k}n - 1$ for any $n \in \mathbb{N}$ for which $r_n \in \mathbb{N}$; we shall assume that such n exist (hence, infinitely many exist), and refer only to such indices.

For all $x \in \operatorname{Irr}_k(n - r_n k) = \operatorname{Irr}_k(k + \gamma n)$, recall that we denoted $\phi_k(x) = (a, b)$ with $b \in \mathbb{Z}_q^{\gamma n}$ in $(0, k-1)_q$- RLL. We shall build a reconstruction code in the descendant cones of only such x, which we denote C_γ.

Lemma 17 There exists a system $\mathcal{S} \subseteq (0,k-1)_q$-RLL and $\theta \in \left(\frac{1}{2},1\right)$ such that

$$\operatorname{cap}(\mathcal{S}) = \lim_{l\to\infty} \frac{1}{l}\log_q\left|\mathcal{S}\cap\mathbb{Z}_q^l\right| = \operatorname{cap}((0,k-1)_q\text{-}\mathrm{RLL})$$

and for all $b\in\mathcal{S}$ it holds that $\mathrm{wt}_H(b) \geqslant \theta|b|$.

Proof: Let $G_q(k-1)$ be the strongly connected deterministic digraph representing the $(0,k-1)_q$-RLL system, seen in Figure 1, whose adjacency matrix is

$$T_q(k-1) = \begin{pmatrix} q-1 & 1 & 0 & \cdots & 0 \\ q-1 & 0 & 1 & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ q-1 & 0 & \cdots & 0 & 1 \\ q-1 & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

As is well known for the case of $q=2$ (see, e.g., [17], [45]), its characteristic polynomial is

$$p_q^{(k-1)}(x) = x^k - (q-1)\sum_{j=0}^{k-1} x^j = \frac{x^{k+1}-qx^k+(q-1)}{x-1},$$

hence the Perron eigenvalue λ of $T_q(k-1)$ is the unique positive root of $\hat{p}_q^{(k-1)}(x) = x^{k+1}-qx^k+(q-1)$ greater than 1 (in fact, $\lambda\in(q-1,q)$, which can readily be confirmed either using elementary calculus or by information-theoretic methods, since $(\mathbb{Z}_q\setminus\{0\})^* \subseteq (0,k-1)_q$-RLL $\subseteq \mathbb{Z}_q^*$).

Further, $T_q(k-1)$ has positive right- and left-eigenvectors associated with λ, which we denote $\bar{v},\bar{w}$ respectively; specifically,

$$\bar{v} = \Bigg(1, \lambda-(q-1), \dots, \lambda^{j-1}-(q-1)\sum_{i=0}^{j-2}\lambda^i, \dots, \\ \dots\lambda^{k-1}-(q-1)\sum_{i=0}^{k-2}\lambda^i\Bigg),$$

$$\bar{w} = \left(\lambda^{k-1},\lambda^{k-2},\dots,\lambda^{k-j},\dots,1\right).$$

and we may verify that

$$v_k = \lambda^{k-1}-(q-1)\sum_{i=0}^{k-2}\lambda^i = \frac{1}{\lambda}\left[\lambda^k-(q-1)\sum_{j=1}^{k-1}\lambda^j\right] \\ = \frac{q-1}{\lambda} > 0$$

and $v_j = \frac{v_{j+1}+(q-1)}{\lambda}$, hence every entry of $\bar{v}$ is indeed positive.

Denoting $q_{i,j} = (T_q(k-1))_{i,j}\cdot\frac{v_j}{\lambda v_i}$, it follows (see, e.g., [37][Sec. 3.5]) that $Q = (q_{i,j})_{1\leqslant i,j\leqslant k}$ is stochastic, and represents a transition matrix of a stationary Markov chain $\mathcal{P}$ on $G_q(k-1)$ (a probability measure on its edges set $E_q(k-1)$) satisfying $H(\mathcal{P}) = \log_q\lambda = \operatorname{cap}((0,k-1)_q\text{-}\mathrm{RLL})$. Further, the stationary distribution of the Markov chain, i.e., a positive $\bar{\pi} = (\pi_1,\dots,\pi_k)$ such that $\sum_{j=1}^k \pi_j = 1$ and $\bar{\pi}^TQ = \bar{\pi}^T$, is given by $\pi_j = \frac{\hat{\pi}_j}{\sum_{i=1}^k\hat{\pi}_i}$, where $\hat{\pi}$ is defined by $\hat{\pi}_j = w_jv_j$. It holds for all j that π_j is the sum of probabilities $\sum\mathcal{P}(e)$ of edges terminating at the j'th node.

Note, then, that

$$\begin{aligned}\sum_{i=1}^k\hat{\pi}_i &= \lambda^{k-1}+\sum_{i=2}^k\left[\lambda^{k-1}-(q-1)\frac{\lambda^{k-1}-\lambda^{k-i}}{\lambda-1}\right] \\ &= \lambda^{k-1}\left[1+(k-1)\left(1-\frac{q-1}{\lambda-1}\right)\right]+\frac{q-1}{\lambda-1}\sum_{i=2}^k\lambda^{k-i} \\ &= \lambda^{k-1}\left[k-(k-1)\frac{q-1}{\lambda-1}\right]+\frac{q-1}{\lambda-1}\sum_{j=0}^{k-2}\lambda^j \\ &= \lambda^{k-1}\left[k-(k-1)\frac{q-1}{\lambda-1}\right]+\frac{\lambda^k-(q-1)\lambda^{k-1}}{\lambda-1} \\ &= \frac{\lambda^{k-1}}{\lambda-1}[\lambda-k(q-\lambda)]\end{aligned}$$

and in particular $\pi_1 = \frac{\lambda-1}{\lambda-k(q-\lambda)}$. (Incidentally, it follows from $\pi_1\in(0,1)$ that $1<k(q-\lambda)<\lambda$, that is, $q-\frac{q}{k+1}<\lambda<q-\frac{1}{k}$.)

Next, recall that for a given $\epsilon>0$, a $(\mathcal{P},\epsilon)$-*strongly-typical* path in G is a path $\gamma = (e_1,e_2,\dots,e_l)$ (denoted by its edges $\{e_1,e_2,\dots,e_l\}\subseteq E_q(k-1)$) such that each $e\in E_q(k-1)$ appears in the path $l\cdot\tau$ times, for some τ satisfying $|\tau-\mathcal{P}(e)|\leqslant\epsilon$. If we let $\mathcal{S}_\epsilon\subseteq\mathbb{Z}_q^*$ be the system induced by $(\mathcal{P},\frac{\epsilon}{k(q-1)})$-strongly-typical paths, then it is well known that $\operatorname{cap}(\mathcal{S}_\epsilon) = \operatorname{cap}((0,k-1)_q\text{-}\mathrm{RLL})$. Note, for $b\in\mathcal{S}_\epsilon$ of length $|b|=l$, which is generated by the path $\gamma=(e_1,\dots,e_l)$, $\mathrm{wt}_H(b)$ is precisely the number of edges which terminate at the first node; since γ is $(\mathcal{P},\frac{\epsilon}{k(q-1)})$-strongly-typical,

$$\mathrm{wt}_H(b) \geqslant \sum_{\substack{e\text{ terminates}\\\text{at first node}}} l\cdot\left(\mathcal{P}(e)-\frac{\epsilon}{k(q-1)}\right) = l(\pi_1-\epsilon)$$

To conclude the proof, note

$$\begin{aligned}\lambda+k(q-\lambda) &= q+(k-1)(q-\lambda) > q \geqslant 2 \\ &\implies \lambda > 2-k(q-\lambda) \\ &\implies 2(\lambda-1) > \lambda-k(q-\lambda) \implies \pi_1 > \frac{1}{2}\end{aligned}$$

Hence we can take any $0<\epsilon<\pi_1-\frac{1}{2}$, and observe that $\mathcal{S}=\mathcal{S}_\epsilon$, $\theta=\pi_1-\epsilon$ satisfy the proposition. ■

Lemma 17 implies that there exists a subset $S_k\subseteq\mathrm{Irr}_k$ such that $\operatorname{cap}(S_k)=\operatorname{cap}(\mathrm{Irr}_k)$, and for every $x\in S_k$ of length $|x|=k+\gamma n$ we have $m(x)\geqslant\lceil\theta\cdot\gamma n\rceil$. For the rest of this section we only build codes C_γ^n in the descendant cones of roots in S_k. Note, then, that if we denote $m_n=\lceil\theta\cdot\gamma n\rceil$ and $\mathcal{C}_\gamma\triangleq\bigcup C_\gamma^n$, then

$$\begin{aligned}\operatorname{cap}(\mathcal{C}_\gamma) &\geqslant \limsup_{n\to\infty}\frac{1}{n}\log_q\Big[|\mathrm{Irr}_k(k+\gamma n)|\cdot \\ &\qquad\qquad\cdot M(m_n,r_n,d_{N,t}(m_n))\Big] \\ &= \gamma\operatorname{cap}(\mathrm{Irr}_k)+ \\ &\quad+\limsup_{n\to\infty}\frac{1}{n}\log_q M(m_n,r_n,d_{N,t}(m_n))\end{aligned} \tag{2}$$

We evaluate the second addend in the following theorem:

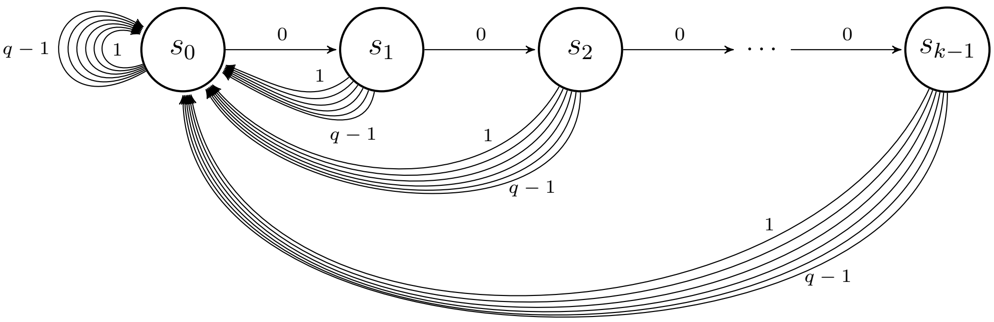

Figure 1. The graph $G_q(k-1)$ generating the $(0,k-1)_q$-RLL system.

Theorem 18 As before, we denote $r_n = \frac{1-\gamma}{k}n - 1$ and $m_n = \lceil \theta \cdot \gamma n \rceil$. Then

$$\lim_{n\to\infty} \frac{1}{n} \log_q M(m_n, r_n, d_{N_n,t_n}(m_n)) = \\ = \frac{\theta\gamma}{\log_2 q} \cdot \mathcal{H}\left(1 + \frac{1-\gamma}{k\theta\gamma}\right)$$

in both of the aforementioned two regimes:

1) Regime I: when $N_n = o(n)$ and $t_n = t$ is fixed.
2) Regime II: when $N_n = 2^{\alpha n}$ and $t_n = \beta n$, if we additionally require $\frac{\alpha^2}{\beta} > 4\theta\gamma$.

Proof:

1) Note, for sufficiently large n, that $N_n < \theta \cdot \gamma n \leqslant m_n$, resulting by Lemma 11 in $d_{N_n,t}(m_n) = t$. We note that $\lim_{n\to\infty} \frac{r_n}{n} = \frac{1-\gamma}{k}$ and $\lim_{n\to\infty} \frac{m_n}{n} = \theta\gamma$, hence by Lemma 10 the claim is proven when t is fixed.

2) By Theorem 15:

$$d_{N_n,t_n}(m_n) \leqslant \max\left\{1, \beta n - \left\lfloor n \frac{\alpha^2 n}{4\lceil \theta \cdot \gamma n \rceil} \right\rfloor\right\} \\ = \max\left\{1, \left\lceil \left(\beta - \frac{\alpha^2 n}{4\lceil \theta \cdot \gamma n\rceil}\right) n \right\rceil\right\}.$$

If $\frac{\alpha^2}{\beta} > 4\theta\gamma$ then for sufficiently large n we have $\beta < \frac{\alpha^2 n}{4\lceil \theta\cdot\gamma n\rceil}$, hence $d_{N_n,t_n}(m_n) = 1$. Since it is fixed, we may now apply the same argument used in the previous part. ■

Going forward, we shall view the lower bound to $\mathrm{cap}(\mathcal{C}_\gamma)$,

$$R(\gamma) \triangleq \gamma \,\mathrm{cap}(\mathrm{Irr}_k) + \frac{\theta\gamma}{\log_2 q} \cdot \mathcal{H}\left(1 + \frac{1-\gamma}{k\theta\gamma}\right),$$

as a function of γ. Before moving on to show that it may be made to exceed $\mathrm{cap}(\mathrm{Irr}_k)$ by a careful choice of γ, we look at the following example.

Example 19 Set $q = k = 2$. Then the Perron eigenvalue of $T_2(1)$ is $\lambda = \frac{1+\sqrt{5}}{2}$, and

$$\mathrm{cap}(\mathrm{Irr}_2) = \log_2(\lambda) = \log_2\left(\frac{1+\sqrt{5}}{2}\right) \approx 0.6942.$$

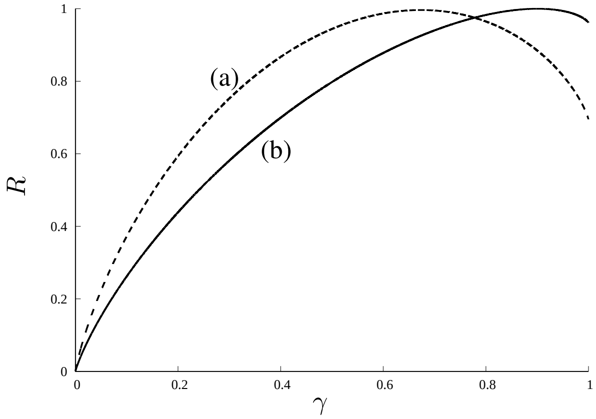

Figure 2. Rate $R(\gamma)$ in the cases (a) $q = k = 2$, $\theta = 0.7236$, and (b) $q = 4$, $k = 2$, $\theta = 0.8273$. The value at $\gamma = 1$ equals $\mathrm{cap}(\mathrm{Irr}_k)$.

In addition, any θ which is less than $\pi_1 = \frac{1}{2}\left(1 + \frac{1}{\sqrt{5}}\right) \approx 0.7236$ satisfies Lemma 17.

Alternatively, we may set $q = 4$ (for the special case of DNA) and duplication-length $k = 2$. Now the Perron eigenvalue of $T_4(1)$ is given by $\lambda = \frac{3+\sqrt{21}}{2}$, hence

$$\mathrm{cap}(\mathrm{Irr}_2) = \log_4(\lambda) = \log_4\left(\frac{3+\sqrt{21}}{2}\right) \approx 0.9613.$$

Further, we may choose any θ which is less than $\pi_1 = \frac{1}{2}\left(1 + \sqrt{\frac{3}{7}}\right) \approx 0.8273$.

$R(\gamma)$ is shown for both cases in Figure 2, under the assumptions of asymptotic regime made in Theorem 18. The figure demonstrates that the capacity of reconstruction codes (bounded from below by the maximum of the curve) is greater than $\mathrm{cap}(\mathrm{Irr}_k)$. □

We now attempt to maximize $R(\gamma)$ by a proper choice of $\gamma \in (0,1)$. Analysis of $R(\gamma)$ is simpler using the following change of variable:

Definition 20 Define $x : (0,1) \to (0,\infty)$ by $x(\gamma) \triangleq \frac{1-\gamma}{\gamma}$.

We observe that $x(\gamma)$ is a decreasing diffeomorphism, and $\gamma = \frac{1}{1+x(\gamma)}$.

Lemma 21 One has

$$R(\gamma) = \gamma\,\mathrm{cap}(\mathrm{Irr}_k) + \theta\gamma\left[\left(1+\frac{x(\gamma)}{k\theta}\right)\log_q\left(1+\frac{x(\gamma)}{k\theta}\right) - \frac{x(\gamma)}{k\theta}\log_q\left(\frac{x(\gamma)}{k\theta}\right)\right]$$

Proof: We observe that for all $x > 0$, $\log\bigl(1+\frac{1}{x}\bigr) = \log\bigl(\frac{x+1}{x}\bigr) = \log(x+1) - \log x$; in particular

$$\log_q\left(1+\frac{k\theta\gamma}{1-\gamma}\right) = \log_q\left(1+\frac{1-\gamma}{k\theta\gamma}\right) - \log_q\left(\frac{1-\gamma}{k\theta\gamma}\right)$$

Hence,

$$\begin{aligned}
R(\gamma) =& \gamma\,\mathrm{cap}(\mathrm{Irr}_k) + \frac{\theta\gamma}{\log_2 q}\cdot\mathcal{H}\left(1+\frac{1-\gamma}{k\theta\gamma}\right)\\
=& \gamma\,\mathrm{cap}(\mathrm{Irr}_k) + \theta\gamma\log_q\left(1+\frac{1-\gamma}{k\theta\gamma}\right)\\
&+ \frac{1-\gamma}{k}\log_q\left(1+\frac{k\theta\gamma}{1-\gamma}\right)\\
=& \gamma\,\mathrm{cap}(\mathrm{Irr}_k) + \left(\theta\gamma+\frac{1-\gamma}{k}\right)\log_q\left(1+\frac{1-\gamma}{k\theta\gamma}\right)\\
&- \frac{1-\gamma}{k}\log_q\left(\frac{1-\gamma}{k\theta\gamma}\right)\\
=& \gamma\,\mathrm{cap}(\mathrm{Irr}_k) + \theta\gamma\left[\left(1+\frac{1-\gamma}{k\theta\gamma}\right)\log_q\left(1+\frac{1-\gamma}{k\theta\gamma}\right)\right.\\
&\left.- \frac{1-\gamma}{k\theta\gamma}\log_q\left(\frac{1-\gamma}{k\theta\gamma}\right)\right]
\end{aligned}$$

■

We can now show that there always exists a choice of γ for which we get $R(C_\gamma^n) > \mathrm{cap}(\mathrm{Irr}_k)$:

Theorem 22 $\max_{\gamma\in(0,1)} R(\gamma) > \mathrm{cap}(\mathrm{Irr}_k)$.

Proof: Observe that $R(\gamma)$ is continuously differentiable and satisfies $\lim_{\gamma\to 0} R(\gamma) = 0$, $\lim_{\gamma\to 1} R(\gamma) = \mathrm{cap}(\mathrm{Irr}_k)$. We find $R'(\gamma)$ in Eq. (3); Thus, We can show that $R'(\gamma) = 0$ if and only if

$$q^{-k\,\mathrm{cap}(\mathrm{Irr}_k)} = \left(1+\frac{x(\gamma)}{k\theta}\right)^{k\theta-1}\cdot\frac{x(\gamma)}{k\theta} \tag{4}$$

This equation has a unique solution $x_0 = x(\gamma_0)$, since the RHS is a monotonic increasing function of x, vanishing at $x = 0$ and unbounded as x grows. Moreover, $0 < x_0 < k\theta$, since $k\theta > 1$, hence the RHS is greater than 1 at $x = k\theta$. Thus $R(\gamma)$ has a unique local extremum in $(0,1)$.

It now suffices to show that $R(\gamma)$ is concave, hence the extremum is a maximum. Indeed,

$$\begin{aligned}
R''(\gamma) &= \frac{1}{k}\frac{dx}{d\gamma}\cdot\frac{d}{dx}\left[(k\theta-1)\log_q\left(1+\frac{x}{k\theta}\right)\right.\\
&\qquad\left.+\log_q\left(\frac{x}{k\theta}\right)\right]_{x=x(\gamma)}\\
&= \frac{-1}{k\ln(q)\gamma^2}\left[\frac{k\theta-1}{k\theta+x(\gamma)}+\frac{1}{x(\gamma)}\right] < 0
\end{aligned}$$

It follows that $R(\gamma_0) > \lim_{\gamma\to 1} R(\gamma) = \mathrm{cap}(\mathrm{Irr}_k)$. ■

Thus, the main result of this paper is established. In what remains of this section we show that we can bound γ_0 which maximizes $R(\gamma)$, in practice, to any desired level of accuracy. We begin by establishing bounds in the following lemma.

Lemma 23 Let $\gamma_0 \in (0,1)$ be the unique maximum of $R(\gamma)$, and denote $x_0 = x(\gamma_0)$. Then

$$x_0 \geqslant \frac{k\theta}{\bigl(2^\theta q^{\mathrm{cap}(\mathrm{Irr}_k)}\bigr)^k - 1}$$

and

$$\begin{aligned}
x_0 &\leqslant \frac{1}{2}\left[\sqrt{\bigl(1-q^{-\mathrm{cap}(\mathrm{Irr}_k)k}\bigr)^2 + k\theta q^{2-\mathrm{cap}(\mathrm{Irr}_k)k}}\right.\\
&\qquad\left.-\left(1-q^{-\mathrm{cap}(\mathrm{Irr}_k)k}\right)\right]\\
&\leqslant \frac{k\theta q^2}{4\bigl(q^{\mathrm{cap}(\mathrm{Irr}_k)k}-1\bigr)}.
\end{aligned}$$

Proof: For fixed $x \in [0,\infty)$ define $g_x : (0,\infty) \to \mathbb{R}$ by $g_x(y) = y\ln\Bigl(1+\frac{x}{y}\Bigr)$. Then

$$\begin{aligned}
g_x'(y) &= \ln\left(1+\frac{x}{y}\right) + \frac{y}{1+\frac{x}{y}}\cdot\frac{-x}{y^2} = \ln\left(1+\frac{x}{y}\right) - \frac{x}{y+x}\\
&= -\ln\left(1-\frac{x}{x+y}\right) - \frac{x}{y+x}\\
&\geqslant -\left(-\frac{x}{x+y}\right) - \frac{x}{y+x} \geqslant 0.
\end{aligned}$$

Therefore, $f_x(y) = e^{g_x(y)} = \Bigl(1+\frac{x}{y}\Bigr)^y$ satisfies $1 + x = f_x(1) \leqslant f_x(y) = \Bigl(1+\frac{x}{y}\Bigr)^y$ for all $y \geqslant 1$. In our case $k\theta > 1$ and x_0 satisfies Eq. (4), hence

$$q^{-\mathrm{cap}(\mathrm{Irr}_k)k} = \left(1+\frac{x_0}{k\theta}\right)^{k\theta-1}\frac{x_0}{k\theta} \geqslant \frac{1+x_0}{1+\frac{x_0}{k\theta}}\cdot\frac{x_0}{k\theta} = \frac{x_0+x_0^2}{k\theta+x_0}$$

which we simplify to $0 \geqslant x_0^2 + \bigl(1-q^{-\mathrm{cap}(\mathrm{Irr}_k)k}\bigr)x_0 - k\theta q^{-\mathrm{cap}(\mathrm{Irr}_k)k}$. Thus, the first upper bound is proven. For the second, we require only that for $a, b > 0$ it holds that $\sqrt{a+b^2} - b \leqslant \frac{a}{2b}$, which is readily shown by differentiation.

On the other hand, Eq. (4) implies that $x_0 \leqslant k\theta$. Therefore

$$\begin{aligned}
q^{-\mathrm{cap}(\mathrm{Irr}_k)k} &= \left(1+\frac{x_0}{k\theta}\right)^{k\theta-1}\frac{x_0}{k\theta} \leqslant \frac{2^{k\theta}}{1+\frac{x_0}{k\theta}}\cdot\frac{x_0}{k\theta}\\
&\iff k\theta q^{-\mathrm{cap}(\mathrm{Irr}_k)k} \leqslant \left(2^{k\theta} - q^{-\mathrm{cap}(\mathrm{Irr}_k)k}\right)x_0
\end{aligned}$$

which proves the lower bound. ■

Next, we show that we may tighten the bounds we derived in the previous lemma.

Lemma 24 Let $x_0 > 0$ be the unique solution to Eq. (4), and denote $z_0 = \frac{x_0}{k\theta}$. If $\underline{z} \leqslant z_0 \leqslant \overline{z}$ then $F(\underline{z}) \leqslant z_0 \leqslant F(\overline{z})$, where

$$F(z) \triangleq \frac{q^{-\mathrm{cap}(\mathrm{Irr}_k)k}}{\left(1+\frac{q^{-\mathrm{cap}(\mathrm{Irr}_k)k}}{(1+z)^{k\theta-1}}\right)^{k\theta-1}}.$$

Proof: By assumption we have $q^{-\mathrm{cap}(\mathrm{Irr}_k)k} = (1+z_0)^{k\theta-1}\cdot z_0$, hence $q^{-\mathrm{cap}(\mathrm{Irr}_k)k} \leqslant (1+\overline{z})^{k\theta-1}\cdot z_0$,

$$\begin{aligned}R'(\gamma) &= \mathrm{cap}(\mathrm{Irr}_k) + \frac{dx}{d\gamma}\cdot\frac{d}{dx}\left[\frac{\theta}{1+x}\left(\left(1+\frac{x}{k\theta}\right)\log_q\left(1+\frac{x}{k\theta}\right)-\frac{x}{k\theta}\log_q\left(\frac{x}{k\theta}\right)\right)\right]_{x=x(\gamma)}\\ &= \mathrm{cap}(\mathrm{Irr}_k) - \frac{1}{\gamma^2}\left[\frac{-\theta}{(1+x)^2}\left(\left(1+\frac{x}{k\theta}\right)\log_q\left(1+\frac{x}{k\theta}\right)-\frac{x}{k\theta}\log_q\left(\frac{x}{k\theta}\right)\right)+\frac{\theta}{(1+x)}\cdot\left(\frac{1}{k\theta}\log_q\left(1+\frac{x}{k\theta}\right)-\frac{1}{k\theta}\log_q\left(\frac{x}{k\theta}\right)\right)\right]_{x=x(\gamma)}\\ &= \mathrm{cap}(\mathrm{Irr}_k) + \frac{1}{k}\left[(k\theta-1)\log_q\left(1+\frac{x(\gamma)}{k\theta}\right)+\log_q\left(\frac{x(\gamma)}{k\theta}\right)\right] \end{aligned} \tag{3}$$

implying that $z_0 \geqslant G(\overline{z})$ where $G(z) = \frac{q^{-\mathrm{cap}(\mathrm{Irr}_k)k}}{(1+z)^{k\theta-1}}$. Similarly, $z_0 \leqslant G(\underline{z})$. The proposition now trivially follows for $F(z) = G(G(z))$. ■

Finally, we can show that x_0 may be found by the following limiting process:

Theorem 25 The unique solution to Eq. (4) is given by $x_0 = k\theta \lim_{n\to\infty} F^n(z_1)$, for all $z_1 \in [0,1]$.

Proof: As before, we denote the unique solution $x_0 > 0$, and take $z_0 = \frac{x_0}{k\theta}$.

Note that Lemma 24 implies that $z_0 = F(z_0)$. We will prove that $F : [0,1] \to [0,1]$ is a contraction; that is, for all $z_1, z_2 \in [0,1]$ we have $|F(z_1) - F(z_2)| \leqslant c|z_1 - z_2|$ for some $c < 1$. Indeed, recalling $k\theta > 1$ we find

$$\begin{aligned}F'(z) &= \frac{2^{-2\,\mathrm{cap}(\mathrm{Irr}_k)k}(k\theta-1)^2}{(1+z)^{k\theta}\left(1+\frac{q^{-\mathrm{cap}(\mathrm{Irr}_k)k}}{(1+z)^{k\theta-1}}\right)^{k\theta}}\\ &\leqslant \frac{(k\theta-1)^2}{(2^{2\,\mathrm{cap}(\mathrm{Irr}_k)})^k} \leqslant \frac{(k-1)^2}{2^k} \leqslant \frac{9}{16} < 1,\end{aligned}$$

where the next to last inequality may be directly verified for all small k.

Having done so, we utilize Banach's fixed-point theorem to deduce that F has a unique fixed point (necessarily z_0), and for all $z_1 \in [0,1]$, defining $z_{n+1} = F(z_n)$ we get $\lim_{n\to\infty} z_n = z_0$. ■

We can now suggest a construction for $(N,t,k)_q$-UTR codes achieving better capacity than the error-correcting codes $\mathrm{Irr}_k(n)$ suggested in [19] (provided that one is willing to consider reconstruction codes over unambiguous decoding of any single output).

Construction A We set the alphabet size q, duplication length k. In the case that our application falls within Regime I, we also set a fixed decoding-delay t, and restrict the ambiguity N_n to be sub-linear in n. (with the necessary adjustments, this construction also applies for Regime II.)

- Start by finding the Perron eigenvalue λ of $T_q(k-1)$, and $\pi_1 = \frac{\lambda-1}{\lambda-k(q-\lambda)}$, as in the proof of Lemma 17. Set some $\theta < \pi_1$.
- The upper and lower bounds on x_0 from Lemma 23 can be made tighter by a repetitive application of $F(\cdot)$ from Lemma 24; Theorem 25 guarantees that the bounds–hence the acceptable error–can be made as tight as desired for our application.
- With $\gamma_0 = \frac{1}{1+x_0}$ we may find $r_n = \frac{1-\gamma_0}{k}n - 1$, and we note that a capacity-achieving subset of $\mathrm{Irr}_k(n - r_n k) = \mathrm{Irr}_k(k + \gamma n)$ has weight $m(x) \geqslant m_n = \lceil \theta\cdot\gamma n\rceil$.
- Within $D_k^{r_n}(x)$ of just such irreducible sequences x we may utilize any construction of codes for the Manhattan metric over $\Delta_{r_n}^{m_n}$ with minimal distance t, if it produces codes of size sufficiently close to $M(m_n, r_n, t)$. For practical applications, [28, Sec. IV-A] showed that if m_n is a prime power, then by [5] there exist such codes of size $\left|\Delta_{r_n}^{m_n}\right| / \frac{m_n^t - 1}{m_n - 1}$ (which improves on the Gilbert-Varshamov bound, and is sufficiently tight to achieve the same result as in Theorem 18).

□

Note that we do not establish that Construction A produces a system of codes of capacity 1, rather only greater than $\mathrm{cap}(\mathrm{Irr}_k)$. To conclude this section, we also present a non-constructive argument proving the existence of a system of reconstruction codes with capacity 1 by an application of the Gilbert-Varshamov bound.

Recall that in the proof of Theorem 18 we have shown that the minimal distance, $d_{N_n,t_n}(m_n)$ was bounded. In particular, in the case of interest Regime I, we used the fact that $m_n = \Theta(n)$; This does not, in general, hold for $m(R_k(y))$ for all $y \in \mathbb{Z}_q^n$.

However, if we show that to be the case for a sufficiently large subset $S^n \subseteq \mathbb{Z}_q^n$, then we may note the following: by [28, Lem. 1] the size of ball in the $d_k(\cdot,\cdot)$ metric of radius d in the descendant cone of $x \in \mathrm{Irr}_k$, where $m(x) \geqslant d$, is

$$\begin{aligned}&\sum_{j=0}^{d}\binom{m(x)}{j}\binom{d}{j}\binom{d+m(x)-j}{d}\\ &\quad\leqslant (d+1)\cdot\binom{m(x)}{d}\binom{d}{\lfloor d/2\rfloor}\binom{d+m(x)}{d}\\ &\quad= O(m(x)^d) = O(n^d)\end{aligned}$$

It would follow that a code of size $\frac{|S^n|}{O(n^d)}$ exists (and, again, the capacity of these codes will be $\mathrm{cap}(S^n)$).

It now suffices to show that except for a vanishingly small portion of $y \in \mathbb{Z}_q^n$, it holds that $m(R_k(y)) = \Theta(n)$. Indeed, recall that $m(R_k(y)) = \mathrm{wt}_H(\mu(b)) = \mathrm{wt}_H(b)$, where $\phi_k(y) = (a,b)$, $b \in \mathbb{Z}_q^{n-k}$. Then, for any real $0 < \xi < 1 - \frac{1}{q}$,

$$\frac{\left|\left\{b \in \mathbb{Z}_q^{n-k} \;\middle|\; \mathrm{wt}_H(b) \leqslant \xi(n-k)\right\}\right|}{q^{n-k}} \leqslant q^{(n-k)(H_q(\xi)-1)},$$

where $H_q(\cdot)$ is the q-ary entropy function,

$$H_q(\xi) \triangleq -\xi\log_q\xi - (1-\xi)\log_q(1-\xi) + \xi\log_q(q-1),$$

and where we used a standard bounding of the size on the Hamming ball, e.g., see [39, Lemma 4.7].

A. Comparison to recent results

Before we finish, we note here that the last argument also shows via the GV bound that error-correcting codes for a fixed number of tandem-duplications achieve capacity 1. Indeed, after the submission of this manuscript [27], [29] were made available, wherein bounds on the optimal size of such error-correcting codes were presented; these bounds show that the redundancy required to correct a fixed number of tandem-duplications is logarithmic in n.

More specifically, both works showed (see [27, Thm. 4], [29, Lem. 6]) that there exist codes $C^n \subseteq \mathbb{Z}_q^n$ that correct up to t tandem-duplications, for a fixed $t \in \mathbb{N}$, satisfying

$$\frac{q^n}{n^t}\left(\frac{q}{q-1}\right)^t \lesssim |C^n|$$

(where we say that $a_n \lesssim b_n$ if $\limsup \frac{a_n}{b_n} \leqslant 1$). They also showed that the optimal size was $\Theta\left(\frac{q^n}{n^t}\right)$. Finally, [27, Lem. 3] demonstrated that C^n can be assumed w.l.o.g. to only contain sequences which roots satisfy $m(x) = \Theta(n)$.

We note that error-correcting codes for t tandem-duplications have minimal $d_k(\cdot,\cdot)$ distance $t+1$; In comparison, then, we have showed that $(N,t,k)_q$-UTR codes, where t is fixed and $N = o(n)$, have minimal distance t (when restricted to descendant cones of irreducible words with $m(x) = \Theta(n)$). The observations above imply that codes designed in the aforementioned works for correcting $t-1$ tandem-duplications, of size $\gtrsim \frac{q^n}{n^{t-1}}\left(\frac{q}{q-1}\right)^{t-1}$, are $(N,t,k)_q$-UTR codes. Importantly, this validates the hypothesis that reconstruction codes for data storage in the DNA of living organisms offer greater data-density than error-correcting codes. Namely, in comparison to the $t\log(n) + O(1)$ redundancy achieved by optimal error-correcting codes in [27], [29], $(N,t,k)_q$-UTR codes achieve redundancy $(t-1)\log(n) + O(1)$.

Finally, we also note for completeness that our results in Regime II, albeit less applicable in practice, are unique to this work.

V. Conclusion

We have proposed that reconstruction codes can be applied to data-storage in the DNA of living organisms, due to the channel's inherent property of data replication.

We have showed, under the assumption of uniform tandem-duplication noise, that any reconstruction code is partitioned into error-correcting codes for the Manhattan metric over a simplex, with minimal distances dependent on the reconstruction parameters. We then proved the existence of reconstruction codes with rate 1, and suggested a construction of a family of codes, which relies on constructions of codes for the simplex. Via Theorem 25, we showed that we can bound the parameters required for code-design in any real application, to any degree of accuracy.

We believe that further research should examine explicit code constructions on the simplex; specifically, encoding and decoding algorithms for sufficiently large codes haven't yet been developed; in addition, only specific asymptotic regimes have been explored, and a gap still exists between lower an upper bounds on the size of non-linear codes. It is also desirable to examine the problem under broader noise models, such as bounded tandem-duplication,interspersed-duplication (perhaps complemented), as well as combinations of multiple error models.

Acknowledgments

The authors gratefully acknowledge the two anonymous reviewers and associate editor, whose insight and suggestions helped shape this paper.

References

[1] J. Acharya, H. Das, O. Milenkovic, A. Orlitsky, and S. Pan, "String reconstruction from substring compositions," *SIAM J. Discrete Math.*, vol. 29, no. 3, pp. 1340–1371, 2015.

[2] N. Alon, J. Bruck, F. Farnoud, and S. Jain, "Duplication distance to the root for binary sequences," *IEEE Trans. on Inform. Theory*, vol. 63, no. 12, pp. 7793–7803, Dec. 2017.

[3] M. Arita and Y. Ohashi, "Secret signatures inside genomic DNA," *Biotechnology Progress*, vol. 20, no. 5, pp. 1605–1607, 2004.

[4] F. Balado, "Capacity of DNA data embedding under substitution mutations," *IEEE Trans. on Inform. Theory*, vol. 59, no. 2, pp. 928–941, Feb. 2013.

[5] R. C. Bose and S. Chowla, "Theorems in the additive theory of numbers," *Commentarii Mathematici Helvetici*, vol. 37, no. 1, pp. 141–147, Dec 1962.

[6] Y. Cassuto and M. Blaum, "Codes for symbol-pair read channels," *IEEE Transactions on Information Theory*, vol. 57, no. 12, pp. 8011–8020, Dec. 2011.

[7] Y. M. Chee, H. M. Kiah, A. Vardy, V. K. Vu, and E. Yaakobi, "Coding for racetrack memories," *IEEE Trans. on Inform. Theory*, vol. 64, no. 11, pp. 7094–7112, Nov. 2018.

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

[9] C. T. Clelland, V. Risca, and C. Bancroft, "Hiding messages in DNA microdots," *Nature*, vol. 399, no. 6736, pp. 533–534, 1999.

[10] O. Elishco, F. Farnoud, M. Schwartz, and J. Bruck, "The capacity of some Pólya string models," in *Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT2016), Barcelona, Spain*, Jul. 2016, pp. 270–274.

[11] F. Farnoud, M. Schwartz, and J. Bruck, "A stochastic model for genomic interspersed duplication," in *Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT2015), Hong Kong, China*, Jun. 2015, pp. 904–908.

[12] ——, "The capacity of string-duplication systems," *IEEE Trans. on Inform. Theory*, vol. 62, no. 2, pp. 811–824, Feb. 2016.

[13] ——, "Estimation of duplication history under a stochastic model for tandem repeats," *BMC Bioinformatics*, vol. 20, no. 1, pp. 64–74, Feb. 2019.

[14] R. Gabrys, H. M. Kiah, and O. Milenkovic, "Asymmetric Lee distance codes for DNA-based storage," *IEEE Trans. on Inform. Theory*, vol. 63, no. 8, pp. 4982–4995, Aug. 2017.

[15] R. Gabrys, E. Yaakobi, and O. Milenkovic, "Codes in the Damerau distance for deletion and adjacent transposition correction," *IEEE Trans. on Inform. Theory*, vol. 64, no. 4, pp. 2550–2570, Apr. 2018.

[16] D. Heider and A. Barnekow, "DNA-based watermarks using the DNA-Crypt algorithm," *BMC Bioinformatics*, vol. 8, no. 1, pp. 176–185, May 2007.

[17] T. D. Howell, "Statistical properties of selected recording codes," *IBM Journal of Research and Development*, vol. 33, no. 1, pp. 60–73, Jan. 1989.

[18] S. Jain, F. Farnoud, and J. Bruck, "Capacity and expressiveness of genomic tandem duplication," *IEEE Trans. on Inform. Theory*, vol. 63, no. 10, pp. 6129–6138, Oct. 2017.

[19] S. Jain, F. Farnoud, M. Schwartz, and J. Bruck, "Duplication-correcting codes for data storage in the DNA of living organisms," *IEEE Trans. on Inform. Theory*, vol. 63, no. 8, pp. 4996–5010, Aug. 2017.

[20] ——, "Noise and uncertainty in string-duplication systems," in *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT2017), Aachen, Germany*, Jun. 2017, pp. 3120–3124.

[21] D. C. Jupiter, T. A. Ficht, J. Samuel, Q.-M. Qin, and P. de Figueiredo, "DNA watermarking of infectious agents: Progress and prospects," *PLoS Pathog*, vol. 6, no. 6, p. e1000950, 2010.

[22] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," *IEEE Trans. on Inform. Theory*, vol. 62, no. 6, pp. 3125–3146, Jun. 2016.

[23] E. Konstantinova, "Reconstruction of permutations distorted by single reversal errors," *Discrete Appl. Math.*, vol. 155, pp. 2426–2434, 2007.

[24] ——, "On reconstruction of signed permutations distorted by reversal errors," *Discrete Math.*, vol. 308, pp. 974–984, 2008.

[25] E. Konstantinova, V. Levenshtein, and J. Siemons, "Reconstruction of permutations distorted by single transposition errors," *arXiv preprint arXiv:math/0702191*, 2007.

[26] M. Kovačević and V. Y. F. Tan, "Improved bounds on Sidon sets via lattice packings of simplices," *SIAM J. Discrete Math.*, vol. 31, no. 3, pp. 2269–2278, 2017.

[27] ——, "Asymptotically optimal codes correcting fixed-length duplication errors in DNA storage systems," *IEEE Communications Letters*, vol. 22, no. 11, pp. 2194–2197, Nov. 2018.

[28] ——, "Codes in the space of multisets–coding for permutation channels with impairments," *IEEE Trans. on Inform. Theory*, vol. 64, no. 7, pp. 5156–5169, Jul. 2018.

[29] A. Lenz, N. Jünger, and A. Wachter-Zeh, "Bounds and constructions for multi-symbol duplication error correcting codes," *arXiv preprint arXiv:1807.02874*, 2018.

[30] A. Lenz, A. Wachter-Zeh, and E. Yaakobi, "Duplication-correcting codes," *Designs, Codes and Cryptography*, vol. 87, no. 2, pp. 277–298, Mar. 2019.

[31] P. Leupold, C. Martín-Vide, and V. Mitrana, "Uniformly bounded duplication languages," *Discrete Appl. Math.*, vol. 146, no. 3, pp. 301–310, 2005.

[32] V. I. Levenshtein, E. Konstantinova, E. Konstantinov, and S. Molodtsov, "Reconstruction of a graph from 2-vicinities of its vertices," *Discrete Appl. Math.*, vol. 156, pp. 1399–1406, 2008.

[33] V. I. Levenshtein and J. Siemons, "Error graphs and the reconstruction of elements in graphs," *J. Combin. Theory Ser. A*, vol. 116, pp. 795–815, 2009.

[34] V. I. Levenshtein, "Efficient reconstruction of sequences," *IEEE Trans. on Inform. Theory*, vol. 47, no. 1, pp. 2–22, Jan. 2001.

[35] M. Liss, D. Daubert, K. Brunner, K. Kliche, U. Hammes, A. Leiherer, and R. Wagner, "Embedding permanent watermarks in synthetic genes," *PLoS ONE*, vol. 7, no. 8, p. e42465, 2012.

[36] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. North-Holland, 1978.

[37] B. H. Marcus, R. M. Roth, and P. H. Siegel, "An introduction to coding for constrained systems," Oct. 2001, unpublished Lecture Notes. [Online]. Available: www.math.ubc.ca/~marcus/Handbook

[38] N. Raviv, M. Schwartz, and E. Yaakobi, "Rank-modulation codes for DNA storage with shotgun sequencing," *IEEE Trans. on Inform. Theory*, vol. 65, no. 1, pp. 50–64, Jan. 2019.

[39] R. M. Roth, *Introduction to Coding Theory*. Cambridge Univ. Press, 2006.

[40] S. L. Shipman, J. Nivala, J. D. Macklis, and G. M. Church, "CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria," *Nature*, vol. 547, p. 345, Jul. 2017.

[41] I. Shomorony, T. A. Courtade, and D. Tse, "Fundamental limits of genome assembly under an adversarial erasure model," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 2, no. 2, pp. 199–208, Dec. 2016.

[42] P. C. Wong, K. k. Wong, and H. Foote, "Organic data memory using the DNA approach," *Commun. ACM*, vol. 46, no. 1, pp. 95–98, Jan. 2003.

[43] E. Yaakobi and J. Bruck, "On the uncertainty of information retrieval in associative memories," *IEEE Trans. on Inform. Theory*, vol. 65, no. 4, pp. 2155–2165, Apr. 2019.

[44] E. Yaakobi, J. Bruck, and P. H. Siegel, "Constructions and decoding of cyclic codes over b-symbol read channels," *IEEE Trans. on Inform. Theory*, vol. 62, no. 4, pp. 1541–1551, Apr. 2016.

[45] E. Zehavi and J. K. Wolf, "On runlength codes," *IEEE Trans. on Inform. Theory*, vol. 34, no. 1, pp. 45–54, Jan. 1988.

Yonatan Yehezkeally (S'12) is a graduate student at the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel. His research interests include coding for DNA storage, combinatorial structures, algebraic coding and finite group theory.

Yonatan received the B.Sc. (*magna cum laude*) and M.Sc. (*summa cum laude*) degrees from Ben-Gurion University of the Negev in 2014 and 2017 respectively, from the department of Mathematics and the department of Electrical and Computer Engineering.

Moshe Schwartz (M'03–SM'10) is an associate professor at the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel. His research interests include algebraic coding, combinatorial structures, and digital sequences.

Prof. Schwartz received the B.A. (*summa cum laude*), M.Sc., and Ph.D. degrees from the Technion – Israel Institute of Technology, Haifa, Israel, in 1997, 1998, and 2004 respectively, all from the Computer Science Department. He was a Fulbright post-doctoral researcher in the Department of Electrical and Computer Engineering, University of California San Diego, and a post-doctoral researcher in the Department of Electrical Engineering, California Institute of Technology. While on sabbatical 2012–2014, he was a visiting scientist at the Massachusetts Institute of Technology (MIT).

Prof. Schwartz received the 2009 IEEE Communications Society Best Paper Award in Signal Processing and Coding for Data Storage.