Uncertainty of Reconstructing Multiple Messages from Uniform-Tandem-Duplication Noise

Yonatan Yehezkeally and **Moshe Schwartz**
School of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer Sheva 8410501, Israel
{yonatany@post,schwartz@ee}.bgu.ac.il

***Abstract*—We propose a list-decoding scheme for reconstruction codes in the context of uniform-tandem-duplication noise, which can be viewed as an application of the associative memory model to this setting. We find the uncertainty associated with $m > 2$ strings (where a previous paper considered $m = 2$) in asymptotic terms, where code-words are taken from a typical set of strings, consisting a growing fraction of the space size, converging to 1. Thus, we find the trade-off between the number of errors, the acceptable list size and the resulting uncertainty, which corresponds to the required number of distinct retrieved outputs for successful reconstruction. It is therefore seen that by accepting list-decoding one may decrease the required number of reads.**

I. Introduction

With recent improvements in DNA sequencing and synthesis technologies, and the advent of CRISPR/Cas gene editing technique [21], the case for DNA as a data-storage medium, specifically *in-vivo*, is now stronger than ever before. It offers a long-lasting and high-density alternative to current storage media, particularly for archival purposes [4]. Moreover, due to medical necessities, the technology required for data retrieval from DNA is highly unlikely to become obsolete, which as recent history shows, cannot be said of concurrent alternatives.

In-vivo DNA storage has somewhat lower data density than in-vitro storage, but it provides a reliable and cost-effective propagation via replication, in addition to some protection to stored data. It also has applications including watermarking genetically modified organisms [1], [7], [17] or research material [10], [24] and concealing sensitive information [5]. However, mutations introduce a diverse set of potential errors, including symbol- or burst-substitution/insertions/deletion, and duplication (including tandem- and interspersed-duplication).

The effects of duplication errors, specifically, were studied in a number of recent works including [8], [9], [11]–[15], [18], [20], [22], [23] among others. These works provided some implicit and explicit constructions for uniform-tandem-duplication codes, as well as some bounds. In [27] the authors then argued that a classical error-correction coding approach is sub-optimal for the application, as it does not take advantage of the cost-effective data replication offered inherently by the medium of in-vivo DNA; instead, it was shown that reframing the problem as a *reconstruction* scheme [16] reduces the redundancy required for any fixed number of duplication errors. In this setting, several (distinct) noisy channel outputs are assumed to be available to the decoder. Since its introduction, several applications of the reconstruction problem to storage technologies were found [2], [3], [25], [26]. Of these, [25] in particular extended the reconstruction model to *associative memory*, where one retrieves the set of all entries (or code-words) *associated* with every element of a given set. For a given size of entry set, the maximal number of entries being possibly associated with all of them was dubbed the *uncertainty* of the memory.

Study of this extended model for in-vivo DNA data storage is motivated by a list-decoding reconstruction scheme, whereby tolerance for decoding a list of possible inputs, given multiple channel outputs, enables reducing the number of required outputs for reconstruction.

This paper focuses on uniform tandem-duplication noise; i.e., we assume throughout that the length of duplication window is fixed. In practical applications, a more complex model where that length is permitted to belong to some set, or perhaps is simply bounded, is more realistic; however, we focus on this model as a step towards that end. Our main goal is to analyze the uncertainty associated with a typical set of strings (consisting of most strings in Σ^n, a definition which is made precise in Lemma 2) as a function of the acceptable list size m, where the number of tandem repeats t which channel outputs undergo is fixed.

The paper is organized as follows: In Section II we describe the main contribution of this paper and put it in context of related works. In Section III we present notations and definitions, then in Section IV we find the uncertainty of the aforementioned typical set in asymptotic form, and develop an efficient decoding scheme.

Throughout the paper a few proofs have been omitted, due to space restrictions; The reader is referred to the arXiv (preprint arXiv:2001.07047) for a complete version.

II. Related works and main contribution

Associative memory was discussed in [25], where items are retrieved by association with other items; the human mind seems to operate in this fashion, one concept bringing up memories of other, related, concepts or events. The more items

This work was supported in part by ISF grant no. 270/18.

one considers together, the smaller the set of items associated with all of them. Giving a precise definition to that notion, one defines the uncertainty of an associative memory as the largest possible size of set $N(m)$ whose members are associated with all elements of an m-subset of the memory code-book.

This model is a generalization of the reconstruction problem posed by Levenshtein in [16], wherein a transmission model is assumed with the decoder receiving multiple channel outputs of the same input. N is then the largest size of intersection of balls of radius t about two distinct code-words, where at most t errors are assumed to have occurred in each transmission; if $N+1$ outputs are available to the decoder, the correct input can be deduced.

This can be viewed as a reduction of the associative memory model to the case of $m=2$, allowing a precise reconstruction of the unique $(m-1=1)$ input. When $m>2$, the decoder seeing $N(m)+1$ channel outputs can only unambiguously infer which list of $l<m$ code-words contains the correct input; thus, a list-decoding model is suggested.

In [27] the authors studied the reconstruction problem for uniform-tandem-duplication noise, which is applicable to in-vivo DNA data storage. An uncertainty which is sub-linear in the message length was assumed (as it represents the number of distinct reads required for decoding), and it was shown that the redundancy required for unique reconstruction was $(t-1)\log_q(n)+O(1)$ (compared to the $t\log_q(n)+O(1)$ redundancy required for unique decoding from a single output [11], [13]), where n is the message length, t the number of errors, and q the alphabet size.

In this paper, we apply the associative memory model from [25] (where binary vectors with the Hamming distance were considered) to the setting of uniform-tandem-duplication noise in finite strings, i.e., we consider list-decoding instead of a unique reconstruction.

Our goal is to find the trade-off between the number of tandem-duplication errors, the uncertainty, and the decoded list size. We find the asymptotic behavior, as the message length n grows, of the uncertainty, or required number of reads (more precisely, that number minus one) N, where it is viewed as a function of the list size (plus one) m, and the number of tandem-duplication errors t. Unlike [27], we use an unrestricted code-book, except to a typical subspace, asymptotically achieving the full space size. This is a first step towards a solution for general codes with a given minimum distance, which would incorporate the code redundancy into the trade-off. Such a solution could also be viewed as an extensions of results in [27], where a unique reconstruction ($m=2$) was considered. Our main contribution (see Theorem 14) can informally be summarized in

$$N=\frac{1+o(1)}{(t-\delta-\lceil\log_n(m)\rceil)!}\left(\frac{q-1}{q}n\right)^{t-\delta-\lceil\log_n(m)\rceil},$$

where $\delta\in\{0,1\}$ is a non-decreasing function in m. Thus, such a trade-off is established.

In conclusion, we show that list-decoding is not only theoretically feasible, but may be efficiently performed. This is done using an isometric transform to integer vectors, and by utilizing combination generators; an efficient list-decoding algorithm is developed, given a sufficient number of distinct channel outputs.

III. Preliminaries

The setting of this paper is the set of finite strings Σ^*, over an alphabet Σ which is assumed to be a finite unital ring of size q (e.g., $\mathbb{Z}_q$, or when q is a prime power, $\mathrm{GF}(q)$).

The length of a string $x\in\Sigma^*$ is denoted $|x|$. A tandem-duplication (or tandem repeat) of fixed duplication-window length k (thus, *uniform* tandem-duplication noise) at index i is defined as follows, for $x,y,z\in\Sigma^*$, $|x|=i$ and $|y|=k$:

$$\mathcal{T}_i(xyz)\triangleq xyyz.$$

Thus, uniform tandem-duplication noise with duplication-window length k acts only on strings of length $\geqslant k$, which we denote $\Sigma^{\geqslant k}$. In order to simplify our analysis, we assume throughout the paper that $k\geqslant 2$.

If $y\in\Sigma^{\geqslant k}$ can be derived from $x\in\Sigma^{\geqslant k}$ by a sequence of tandem repeats, i.e., if there exist $i_1,\ldots,i_t$ such that

$$y=\mathcal{T}_{i_t}(\cdots\mathcal{T}_{i_1}(x)),$$

then y is called a *t-descendant* (or simply *descendant*) of x (vice versa, x is an *ancestor* of y), and we denote $x\overset{t}{\Longrightarrow}y$. We say that x is a 0-descendant of itself. If $t=1$ we denote $x\Longrightarrow y$. Where the number of repeats is unknown or irrelevant, we may denote $x\overset{*}{\Longrightarrow}y$. We define the set of t-descendants of x as

$$D^t(x)\triangleq\left\{y\in\Sigma^*:x\overset{t}{\Longrightarrow}y\right\},$$

and the *descendant cone* of x as

$$D^*(x)\triangleq\left\{y\in\Sigma^*:x\overset{*}{\Longrightarrow}y\right\}=\bigcup_{t=0}^{\infty}D^t(x).$$

If there exists no $z\neq x$ such that $z\overset{*}{\Longrightarrow}x$, we say that x is *irreducible*. The set of irreducible strings of length n is denoted $\mathrm{Irr}(n)$. It can be shown (see, e.g., [9]) that for all $y\in\Sigma^{\geqslant k}$ there exists a unique irreducible x, called the *duplication root* of y and denoted $\mathrm{drt}(y)$, such that $y\in D^*(x)$. This induces a partition of $\Sigma^{\geqslant k}$ into descendant cones; i.e., it induces an equivalence relation, denoted herein $\sim_k$.

A useful tool in studying uniform tandem-duplication noise is the *discrete derivative* ϕ defined for $x\in\Sigma^{\geqslant k}$ as $\phi(x)\triangleq\hat{\phi}(x)\bar{\phi}(x)$, where

$$\hat{\phi}(x)\triangleq x(1),x(2),\ldots,x(k),$$
$$\bar{\phi}(x)\triangleq x(k+1)-x(1),\ldots,x(|x|)-x(|x|-k).$$

As seen, e.g., in [9], ϕ is injective, and if $\bar{\phi}(x)=uv$ for $u,v\in\Sigma^*$, $|u|=i$, then $\bar{\phi}(\mathcal{T}_i(x))=u0^kv$. This was used in [27] to define the function $\psi_x\colon D^*(x)\to\mathbb{N}^{w+1}$ by

$$\psi_x(y)\triangleq(\lfloor u(1)/k\rfloor,\ldots,\lfloor u(w+1)/k\rfloor),$$

if

$$\bar{\phi}(y)=0^{u(1)}a_1 0^{u(2)}\ldots a_w 0^{u(w+1)},$$

where $w = \mathrm{wt}(\bar{\phi}(x))$ and $a_1 \ldots, a_w \in \Sigma \setminus \{0\}$. It was shown that ψ_x is a poset isomorphy, where $D^*(x)$ is ordered with $\overset{*}{\Longrightarrow}$ and $\mathbb{N}^{w+1}$ with the product order.

A metric can be defined on $D^r(x)$ for each r (in particular, but not necessarily, when x is irreducible) by

$$d(y_1, y_2) \triangleq \min\big\{t \in \mathbb{N} : D^t(y_1) \cap D^t(y_2) \neq \emptyset\big\},$$

and it is seen in [9] that this is well defined, in the sense that there does exist such t, for $y_1, y_2 \in D^t(x)$, such that $D^t(y_1) \cap D^t(y_2) \neq \emptyset$.

If we define on $\mathbb{N}^{w+1}$ the 1-norm $\|u\|_1 \triangleq \sum_{i=1}^{w+1} u(i)$ and metric $d_1(u,v) \triangleq \frac{1}{2}\|u - v\|_1$, then ψ_x is also an isometry (see [27]) between $D^r(x)$, for each r, and its image in $\mathbb{N}^{w+1}$, which is the simplex

$$\Delta_r^w \triangleq \big\{u \in \mathbb{N}^{w+1} : \|u\|_1 = r\big\} = \psi_x(D^r(x)).$$

The focus of this paper is to find the uncertainty, after t tandem repeats, as a function of the acceptable list size m. This is made precise by the following definition.

Definition 1 Given $n, t \in \mathbb{N}$ and $x_1, \ldots, x_m \in \Sigma^n$, we define

$$S_t(x_1, \ldots, x_m) \triangleq \bigcap_{i=1}^{m} D^t(x_i).$$

Then, the *uncertainty* associated with a code $C \subseteq \Sigma^n$ is

$$N_t(m, C) \triangleq \max_{\substack{x_1,\ldots,x_m \in C \\ x_i \neq x_j}} |S_t(x_1, \ldots, x_m)|.$$

Correspondingly, for $w, r \in \mathbb{N}$ and $u_1, \ldots, u_m \in \Delta_r^w$ we define

$$\bar{S}_t(u_1, \ldots, u_m) \triangleq \bigcap_{i=1}^{m} \big\{v \in \mathbb{N}^{w+1} : v \geqslant u_i,\ \|v - u_i\|_1 = t\big\};$$
$$\bar{N}_t(m, w, r) \triangleq \max_{u_1,\ldots,u_m \in \Delta_r^w} \big|\bar{S}_t(u_1, \ldots, u_m)\big|.$$

In the next section we describe a typical set of strings in Σ^n, then by ascertaining $\bar{N}_t(m, w, r)$ for that set we find an asymptotic expression (in the string length n) for the uncertainty associated with that set, as a function of m.

IV. Typical set

We observe that the sets introduced in the previous section have many parameters. A complete combinatorial analysis of those would be riddled with pathological extreme cases, tedious, and not enlightening; this is particularly so since these extreme cases occur in a vanishingly small fraction of the space. Since our main goal is an asymptotic analysis, we proceed by eliminating those rare pathological cases, and focus on the common typical ones. In particular, we would like to limit our attention to strings $x \in \Sigma^n$ for which the Hamming weight of $\bar{\phi}(x)$ and the 1-norm of $\psi_{\mathrm{drt}(x)}(x)$, as well as the difference between them, are asymptotically linearly proportional to the string length n. Those strings would form the code which we study. Thus, we start by presenting in the following lemma the code C for which it shall be our goal to find $N_t(m, C)$.

Lemma 2 Define the family of codes

$$\mathrm{Typ}^n \triangleq \left\{x \in \Sigma^n : \begin{smallmatrix} \left|w(x) - \frac{q-1}{q}(n-k)\right| < n^{3/4} \\ \left|r(x) - \frac{q-1}{q(q^k-1)}(n-k)\right| < 2n^{3/4} \end{smallmatrix}\right\},$$

where $w(x) \triangleq \mathrm{wt}_H\big(\bar{\phi}(x)\big)$ and $r(x) \triangleq \big\|\psi_{\mathrm{drt}(x)}(x)\big\|_1$. Then for sufficiently large n:

$$\frac{|\mathrm{Typ}^n|}{|\Sigma^n|} \geqslant 1 - 4e^{-\sqrt{n}/2} \underset{n\to\infty}{\longrightarrow} 1.$$

The proof, appearing in the arXiv version, is omitted here.

We remark that a similar concentration result (for $w(x)$ and $\mathrm{wt}_H(\psi_{\mathrm{drt}(x)}(x))$ instead of $r(x)$) was derived in [11, Lem. 3]; it uses a different approach to the one seen in the arXiv version of this work.

Next, for Typ^n we show that the uncertainty can be calculated by $\bar{N}_t$, which provides an expression we may more easily analyze.

Lemma 3 If $C \subseteq \Sigma^n$ and $x_1, \ldots, x_m \in C$, $x_i \neq x_j$, such that $|S_t(x_1, \ldots, x_m)| = N_t(m, C)$, then there exists $x = \mathrm{drt}(\{x_1, \ldots, x_m\})$, and

$$|S_t(x_1, \ldots, x_m)| = \big|\bar{S}_t(\psi_x(x_1), \ldots, \psi_x(x_m))\big|.$$

Proof: If there exist $x_i \not\sim_k x_j$, then $S_t(x_1, \ldots, x_m) = \emptyset$. Otherwise the claim follows from the isometry ψ_x. ∎

Corollary 4 For $k \geqslant 2$ and sufficiently large n,

$$N_t(m, \mathrm{Typ}^n) = \\ = \max\left\{\bar{N}_t(m, w, r) : \begin{smallmatrix} \left|w - \frac{q-1}{q}(n-k)\right| < n^{3/4} \\ \left|r - \frac{q-1}{q(q^k-1)}(n-k)\right| < 2n^{3/4} \end{smallmatrix}\right\}.$$

The proof, appearing in the arXiv version, is omitted here.

Hence, the quantity one needs to assess is $\bar{N}_t(m, w, r)$. We do that next by exploiting the lattice structure of $\mathbb{N}^{w+1}$, and introducing the connection to supremum height and lower-bound-set size in that lattice.

Lemma 5 Given $u_1, \ldots, u_m \in \Delta_r^w$, denote $u \triangleq \bigvee_{i=1}^m u_i$. Then,

$$\big|\bar{S}_t(u_1, \ldots, u_m)\big| = \begin{cases} 0 & \|u\|_1 > r + t, \\ \binom{w+t+r-\|u\|_1}{w} & \text{otherwise.} \end{cases}$$

Proof: The proposition follows from the lattice structure of $\mathbb{N}^{w+1}$, i.e.,

$$\bar{S}_t(u_1, \ldots, u_m) = \left\{v \in \mathbb{N}^{w+1} : v \geqslant \bigvee_{i=1}^{m} u_i,\ \|v - u_1\|_1 + t\right\}$$

∎

Definition 6 Denote the *minimum supremum height*

$$\sigma(m, w, r) \triangleq \min_{u_1,\ldots,u_m \in \Delta_r^w} \Big\|\bigvee\nolimits_{i=1}^{m} u_i\Big\|_1 - r.$$

Conversely, for $w, r, s \in \mathbb{N}$ and $u \in \Delta^w_{r+s}$, denote the *lower-bounds set* $A_r(u) \triangleq \{v \in \Delta^w_r : v \leqslant u\}$ and the *maximal lower-bounds-set size*

$$\mu(w, r, s) \triangleq \max\big\{|A_r(u)| : u \in \Delta^w_{r+s}\big\}.$$

Corollary 7 $\bar{N}_t(m, w, r) = \binom{w+t-\sigma(m,w,r)}{w}$.

Proof: The proposition follows from Lemma 5. ∎

It is therefore seen that the main task is to find or estimate the minimum supremum height. We next show the duality between $\sigma(m, w, r)$ and $\mu(w, r, s)$, which we shall use to calculate the former.

Lemma 8 Take $w, r, s \in \mathbb{N}$. If $s \geqslant wr$ then

$$\mu(w, r, s) = |\Delta^w_r| = \binom{r+w}{r} \quad \text{and} \quad \sigma(|\Delta^w_r|, w, r) = wr.$$

For $s < wr$ we have $\sigma(\mu(w, r, s), w, r) = s$.

The proof, appearing in the arXiv version, is omitted here.

Corollary 9 If $\mu(w, r, s) < m \leqslant \mu(w, r, s+1)$ then $\sigma(m, w, r) = s + 1$.

Proof: Firstly, since $m \mapsto \sigma(m, w, r)$ is non-decreasing by definition,

$$s = \sigma(\mu(w, r, s), w, r) \leqslant \sigma(m, w, r) \leqslant \\ \leqslant \sigma(\mu(w, r, s+1), w, r) = s + 1.$$

However, if $\sigma(m, w, r) = s$, by finding $u_1, \ldots, u_m \in \Delta^w_r$ with $\|\bigvee_{i=1}^m u_i\|_1 = r + s$ we deduce $\mu(w, r, s) \geqslant m$, in contradiction. ∎

Since we now know that calculating $\mu(w, r, s)$ is sufficient for our purposes, we turn to that task; since our focus is Typ^n, we may do so for the relevant ranges of w, r, where that is simpler.

Lemma 10 For $w, r, s \in \mathbb{N}$ there exists $u \in \Delta^w_{r+s}$ such that $|A_r(u)| = \mu(w, r, s)$ and for all $1 \leqslant i < j \leqslant w + 1$ it holds that $|u(i) - u(j)| < 2$.

Proof: Take $u \in \Delta^w_{r+s}$ satisfying $|A_r(u)| = \mu(w, r, s)$, and assume to the contrary that there exist i, j such that, w.l.o.g., $u(j) \geqslant u(i) + 2$. Denote by u' the vector which agrees on u on all coordinates except $u'(j) = u(j) - 1$ and $u'(i) = u(i) + 1$.

Further, partition $A_r(u)$ and $A_r(u')$ by the projection on all other coordinates. For any matching classes $C, C' \subseteq \Delta^w_r$ in the corresponding partitions, denote by $t(C) = t(C')$ the difference between r and the sum of all coordinates other than i, j; Note that $|C|$ is the number of ways to distribute $t(C)$ balls into two bins with capacities $u(i), u(j)$ (and correspondingly $u'(i), u'(j)$ for $|C'|$), hence

$$\begin{aligned}|C| &= \min\{t(C), u(i)\} - \max\{t(C) - u(j), 0\} + 1 \\ &\leqslant \min\{t(C), u(i) + 1\} - \max\{t(C) - u(j) + 1, 0\} + 1 \\ &= \min\{t(C'), u'(i)\} - \max\{t(C) - u'(j), 0\} + 1 = |C'|,\end{aligned}$$

where the inequality is justified by cases for $t(C)$, and is strict only if $u(i) < t(C) < u(j)$. Thus, the proof is concluded. ∎

Lemma 10 allows us to find $\mu(w, r, s)$ with relative ease; perhaps the most straightforward example of that is a precise calculation for the cases $s = 1, 2$, which we present next; following the examples we conduct a more extensive evaluation, for $s > 2$ and the relevant ranges of w, r.

Example 11 Any vector $u \in \Delta^w_{r+1}$ having $1 + \min\{w, r\}$ positive coordinates has precisely $|A_r(u)| = 1 + \min\{w, r\}$. By Lemma 10 one such vector satisfies $\mu(w, r, 1) = |A_r(u)|$, therefore $\mu(w, r, 1) = 1 + \min\{w, r\}$. □

Example 12 We define an injection $\xi\colon \{v \in \mathbb{N}^{w+1} : v \leqslant u\} \to \mathbb{N}^{w+1}$ by $\xi(v) \triangleq u - v$; then clearly, ξ is distance preserving, and in particular injective. Hence, $\mu(w, r, 2) \leqslant |\Delta^w_2| = \binom{w+2}{2}$. This is achieved with equality when $r + 2 \geqslant 2(w + 1)$, as evidenced by any vector greater than $(2, 2, \ldots, 2)$. The inequality is strict, however, when $r < 2w$.

To examine the remaining cases, note first that increasing any coordinate of u above 2 has no effect on $|A_r(u)|$. Further, we again know by Lemma 10 that $\mu(w, r, 2)$ is achieved when u has the greatest number of positive coordinates, and among such vectors, the greatest number greater than or equal to 2. Now, by counting the number of lower bounds for any such $u \in \Delta^w_{r+2}$ we see that

$$\mu(w, r, 2) = \begin{cases} \binom{w+2}{2}, & r \geqslant 2w; \\ \binom{w+1}{2} + (r - w + 1), & w - 1 \leqslant r < 2w; \\ \binom{r+2}{2}, & r < w - 1. \end{cases}$$

□

As can now be seen, a complete evaluation of $\mu(w, r, s)$ for $s > 2$ is possible using Lemma 10, but it involves application of the inclusion-exclusion principle and its results are not illuminating. We shall see instead that an asymptotic evaluation of $\mu(w, r, s)$ for typical ranges of w, r will suffice. To do so, we note the following proposition.

Lemma 13 Fix t, and take w, r such that $r + t \leqslant w + 1$. For all $s \leqslant t$ it holds that $\mu(w, r, s) = \binom{r+s}{s}$.

Proof: By Lemma 10 we know that $u \in \Delta^w_{r+s}$ achieving $|A_r(u)| = \mu(w, r, s)$ is such that $r + s$ of its coordinates equal 1, and the remaining $w + 1 - r - s$ equal 0. The proposition follows. ∎

We can use what we now know about maximal size of lower-bounds sets to establish the main result of this paper, in the following theorem. Before doing so, we note a consequence of, e.g., Lemma 13, namely that for any string $x \in \mathrm{Typ}^n$, and any $y \in D^t(x)$, it holds that

$$\big|\{x' \in \mathrm{Typ}^n : y \in D^t(x')\}\big| = O(n^t).$$

Hence, we have for $m_n = \omega(n^t)$ that $N_t(m_n, \mathrm{Typ}^n) = o(1)$; it is therefore only interesting to find an asymptotic expression for $N_t(m_n, \mathrm{Typ}^n)$ when $m_n = O(n^t)$.

Theorem 14 Fix t and a sequence $m_n = O(n^t)$. Then

$$N_t(m_n, \mathrm{Typ}^n) = \tfrac{1+o(1)}{(e_t(m_n,n))!}\Big(\tfrac{q-1}{q}n\Big)^{e_t(m_n,n)},$$

where $e_t(m_n, n) = t - \lceil \log_n(m_n) \rceil - \delta(m_n, n)$ and $\delta(m, n) \in \{0, 1\}$ is a non-decreasing function in m.

Proof: Let $s \triangleq \lceil \log_n(m_n) \rceil$.

Recall from Lemma 13 that for $w \geqslant r + t - 1$

$$\mu(w, r, s-1) = \binom{r+s-1}{r} < \frac{(r+s-1)^{s-1}}{(s-1)!},$$

hence for r satisfying $\left| r - \frac{q-1}{q(q^k-1)}(n-k) \right| < 2n^{3/4}$ and sufficiently large n

$$\log_n \mu(w, r, s-1) < s - 1.$$

On the other hand we have

$$\mu(w, r, s+1) = \binom{r+s+1}{r} > \frac{r^{s+1}}{(s+1)!},$$

and therefore, for such r,

$$\begin{aligned}\log_n \mu(w, r, s+1) &> \log_n \left(\frac{1+o(1)}{(s+1)!} \left(\frac{q-1}{q(q^k-1)} n \right)^{s+1} \right) \\ &= s + 1 + o(1).\end{aligned}$$

Since $s - 1 < \log_n(m_n) \leqslant s$ it now follows from Corollary 9, for sufficiently large n (which does not depend on s, i.e., on m_n), and w, r satisfying $\left| w - \frac{q-1}{q}(n-k) \right| < n^{3/4}$ and $\left| r - \frac{q-1}{q(q^k-1)}(n-k) \right| < 2n^{3/4}$, that

$$\sigma(m_n, w, r) = s + \delta(m_n, n, r),$$

where

$$\delta(m_n, n, r) = \begin{cases} 1, & m_n > \binom{r+\lceil \log_n(m_n) \rceil}{r}; \\ 0, & \text{otherwise.} \end{cases}$$

Next, for such n, w, r we have

$$\begin{aligned}&\binom{w + t - \sigma(m_n, w, r)}{w} \\ &\qquad = \tfrac{1+o(1)}{(t-(s+\delta(m_n,n,r)))!}\Big(\tfrac{q-1}{q}n\Big)^{t-(s+\delta(m_n,n,r))}.\end{aligned}$$

It therefore follows from Corollary 4 and Corollary 7 that

$$\begin{aligned}N_t(m_n, \mathrm{Typ}^n) &= \tfrac{1+o(1)}{(t-(s+\delta(m_n,n)))!}\Big(\tfrac{q-1}{q}n\Big)^{t-(s+\delta(m_n,n))} \\ &= \tfrac{1+o(1)}{e_t(m_n,n)!}\Big(\tfrac{q-1}{q}n\Big)^{e_t(m_n,n)},\end{aligned}$$

where $\delta(m_n, n) = 1$ if and only if $\delta(m_n, n, r) = 1$ for all r satisfying the above requirement, and $e_t(m_n, n)$ is as defined in the theorem's statement. ∎

Finally, we note that the process of list-decoding given $N_t(m, \mathrm{Typ}^n) + 1$ distinct strings in Σ^{n+kt}, i.e., finding $x_1, \ldots, x_l \in \mathrm{Typ}^n$, $l < m$, such that these strings lie in $S_t(x_1, \ldots, x_l) \setminus \bigcup_{x \in \mathrm{Typ}^n \setminus \{x_1, \ldots, x_l\}} D^t(x)$, is straightforward:

Algorithm A Denote $N \triangleq N_t(m, \mathrm{Typ}^n)$ and assume as input distinct $y_1, \ldots, y_{N+1} \in \Sigma^{n+kt}$ such that there exists $x \in \mathrm{Typ}^n$ satisfying $y_1, \ldots, y_{N+1} \in D^t(x)$.

1) Apply $\psi_{\mathrm{drt}(y_1)}$ to map them to $v_1, \ldots, v_{N+1} \in \Delta^w_{r+t}$ where $w = \mathrm{wt}\big(\bar{\phi}(\mathrm{drt}(y_1))\big)$ and $r = \big\|\psi_{\mathrm{drt}(y_1)}(y_1)\big\|_1$; note that prior computation of $\mathrm{drt}(y_1)$ is not required to perform this mapping, and that it may be found as a byproduct of finding any v_i.
2) Find $u \triangleq \bigwedge_{i=1}^{N+1} v_i \in \Delta^w_{r'}$ by calculating the minimum over each coordinate.
3) Calculate $A_r(u)$.
4) Return $\psi^{-1}_{\mathrm{drt}(y_1)}(A_r(u))$ as a list.

Theorem 15 Algorithm A operates in $O(n^t) = \mathrm{poly}(N)$ steps, and produces $x_1, \ldots, x_l \in \mathrm{Typ}^n$, $l < m$, such that

$$y_1, \ldots, y_{N+1} \in S_t(x_1, \ldots, x_l) \setminus \bigcup_{\substack{x \in \mathrm{Typ}^n \\ x \notin \{x_1, \ldots, x_l\}}} D^t(x).$$

Proof: First, note that the existence of an ancestor for all $y_1, \ldots, y_{N+1}$ implies that $y_i \in D^*(\mathrm{drt}(y_1))$ for all i. Moreover, note that finding any v_i may be done in $O(n)$ steps (by calculating $\bar{\phi}(y_i)$ and recording lengths of runs of zeros in the process). Any one of these can also produce $\mathrm{drt}(y_1)$. Hence Step 1 concludes in $O(Nn)$ steps.

Step 2 can also be performed in $O(Nw) = O(Nn)$ steps.

Now, note that since an ancestor of all y_i's exists in Σ^n, $r' \geqslant r$. It is hence possible to compute $A_r(u)$. This may be achieved by finding all ways of distributing $r' - r < t$ balls into $w + 1$ bins with capacities $u(j)$, e.g., by utilizing combination generators for all $\binom{w+r'-r}{w}$ combinations, then discarding combination which violate the bin-capacity restriction. Combination generating algorithms exist which generate all combinations in $O\Big(\binom{w+r'-r}{w}\Big) = O(n^{t-1})$ steps (e.g., see [19]), and pruning illegal combinations can be done in $O(w)$ steps each. Step 3 can therefore be performed in $O(n^t)$ steps.

Finally, the pre-image $\psi^{-1}_{\mathrm{drt}(y_1)}(A_r(u))$ is a set of ancestors of $y_1, \ldots, y_{N+1}$, which is a subset Typ^n, and no other element of Typ^n is an ancestor of $y_1, \ldots, y_{N+1}$. We also know that $|A_r(u)| < m$, otherwise a contradiction is reached to the definition of N. Computing $\psi^{-1}_{\mathrm{drt}(y_1)}(A_r(u))$ given $\mathrm{drt}(y_1)$ requires $O(|A_r(u)|w) \leqslant O(mn)$ steps. ∎

ACKNOWLEDGMENTS

The authors gratefully thank Prof. Jehoshua Bruck for his insight and suggestions, which originally turned our attention to the problem explored in this paper.

REFERENCES

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

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

[3] 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.

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

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

[6] J. L. Doob, "Regularity properties of certain families of chance variables," *Transactions of the American Mathematical Society*, vol. 47, no. 3, pp. 455–486, 1940.

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

[8] 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.

[9] 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.

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

[11] M. Kovačević and V. Y. F. Tan, "Asymptotically optimal codes correcting fixed-length duplication errors in DNA storage systems," *IEEE Communications Letters*, vol. 22, no. 11, pp. 2194–2197, Nov. 2018.

[12] M. Kovačević, "Codes correcting all patterns of tandem-duplication errors of maximum length 3," *arXiv preprint arXiv:1911.06561*, 2019.

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

[14] A. Lenz, A. Wachter-Zeh, and E. Yaakobi, "Bounds on codes correcting tandem and palindromic duplications," *arXiv preprint arXiv:1707.00052*, 2017.

[15] ——, "Duplication-correcting codes," *Designs, Codes and Cryptography*, vol. 87, no. 2, pp. 277–298, Mar. 2019.

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

[17] 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.

[18] H. Mahdavifar and A. Vardy, "Asymptotically optimal sticky-insertion-correcting codes with efficient encoding and decoding," in *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT2017), Aachen, Germany*, Jun. 2017, pp. 2683–2687.

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

[20] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Exact reconstruction from insertions in synchronization codes," *IEEE Trans. on Inform. Theory*, vol. 63, no. 4, pp. 2428–2445, Apr. 2017.

[21] 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.

[22] Y. Tang and F. Farnoud (Hassanzadeh), "Error-correcting codes for noisy duplication channels," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2019, pp. 140–146.

[23] Y. Tang, Y. Yehezkeally, M. Schwartz, and F. Farnoud (Hassanzadeh), "Single-error detection and correction for duplication and substitution channels," in *Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT) (ISIT'2019), Paris, France*, Jul. 2019, pp. 300–304.

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

[25] 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.

[26] 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.

[27] Y. Yehezkeally and M. Schwartz, "Reconstruction codes for DNA sequences with uniform Tandem-Duplication errors," *IEEE Trans. on Inform. Theory*, 2019, to appear.