On Codes for the Noisy Substring Channel

Yonatan Yehezkeally* and Nikita Polyanskii*†
*Institute for Communications Engineering,
Technical University of Munich, 80333 Munich, Germany,
{yonatan.yehezkeally,nikita.polianskii}@tum.de
†Center for Computational and Data-Intensive Science and Engineering,
Skolkovo Institute of Science and Technology, 121205 Moscow, Russia

***Abstract*—We consider the problem of coding for the substring channel, in which information strings are observed only through their (multisets of) substrings. Because of applications to DNA-based data storage, due to DNA sequencing techniques, interest in this channel has renewed in recent years. In contrast to existing literature, we consider a noisy channel model, where information is subject to noise *before* its substrings are sampled, motivated by in-vivo storage.**

We study two separate noise models, substitutions or deletions. In both cases, we examine families of codes which may be utilized for error-correction and present combinatorial bounds. Through a generalization of the concept of repeat-free strings, we show that the added required redundancy due to this imperfect observation assumption is sublinear, either when the fraction of errors in the observed substring length is sufficiently small, or when that length is sufficiently long. This suggests that no asymptotic cost in rate is incurred by this channel model in these cases.

I. Introduction

DNA as a medium for data storage offers high density and longevity, far greater than those of electronic media [1]. Among its applications, data storage in DNA may offer a protected medium for long-period data storage [2], [3]. In particular, it has recently been demonstrated that storage in the DNA of living organisms (henceforth, *in-vivo* DNA storage) is now feasible [4]; the envelope of a living cell affords some level of protection to the data, and even offers propagation, through cell replication. Among its varied usages, in-vivo DNA storage allows watermarking genetically modified organisms (GMOs) [5]–[7] to protect intellectual property, or labeling research material [3], [8]. It may even conceal sensitive information, as it may appear indistinguishable from the organism's own genetic information [9].

Similarly to other media, information stored over this medium is subjected to noise due to mutations, creating errors in data, which accumulate over time and replication cycles. Examples of such noise include symbol insertions or deletion, in addition to substitutions (point-mutations); the latter is the focus of the vast majority of classical error-correction research, and the former have also been studied. Interestingly, however, the very methods we currently use to store and later retrieve data from DNA inherently introduce new constraints on information reconstruction. While desired sequences may be synthesized (albeit, while suffering from, e.g., substitution noise), the process of DNA sequencing, i.e., retrieving the DNA sequence of an organism, only observes that sequence as the (likely incomplete) multiset of its substrings (practically, up to a certain substring length) [10]. Thus, information contained in the order of these substrings might be irrevocably lost. As a result of these constraints, conventional and well-developed error-correction approaches cannot simply be applied.

To overcome these effects, one approach in existing literature is to add redundancy in the form of indexing, in order to recover the order of substrings (see, e.g., [11]–[13]). A different approach, potentially more applicable to in-vivo DNA storage, is to add redundancy in the form of constraints on the long information string, such that it can be uniquely reconstructed by knowledge of its substrings of a given length (or range of lengths). The combinatorial problem of recovering a sequence from its substrings has attracted attention in recent years [14]–[21], and coding schemes involving only these substrings (including the incidence frequency of each substring) were studied [10], [11], [22]–[24].

However, works dedicated to overcoming this obstacle, inherent to the technology we use, have predominantly focused on storage outside of living cells (i.e., *in-vitro* DNA storage). Likewise, works focused on error-correction for in-vivo DNA data storage (e.g., [25]–[27]) have disregarded the technical process by which data is to be read. However, in real applications varied distinct noise mechanisms act on stored data concurrently. Hence, in practice, both sets of challenges have to be collectively overcome in order to robustly store information using in-vivo DNA.

The aim of this work is to protect against errors in the information string (caused by mutations over the replication process of cells), when channel outputs are given by the multisets of their substrings, of a predetermined length, rather than entire strings. This models the process of DNA sequencing, once information needs to be read from the medium. We shall study the required redundancy of this model, and devise coding strategies, under the assumption of two different error types: substitution and deletion noise.

This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Grant 801434. Y. Yehezkeally's work is supported by the Alexander von Humboldt-Stiftung/Foundation in cooperation with the Carl Friedrich von Siemens Foundation. N. Polyanskii's work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under Grant No. WA3907/1-1.

The paper is organized as follows. In Section II, we discuss the main contribution of this paper, in context of related works. In Section III we then present necessary notation. Finally, in Section IV-A we study the suggested model with substitution errors, and in Section IV-B with deletion errors. Due to space limitation, some proofs have been sketched or delegated to a preprint of the manuscript [28].

II. Related works and main contribution

Given a string of length n, the problem of reconstructing x from the multiset of (all-, or, in some works, most-) its substrings of a fixed lenght $s \leqslant n$, has been studied in literature. Assuming no errors occur in x prior to sampling of its substrings, the problem of interest is identifying a set of constraints on the information string, equivalent of sufficient, for such reconstruction to be achievable.

It was observed in [14] that under certain circumstances, distinct information strings in which repetitions of s-substrings appear in different positions, exhibit the same multisets of $(s+1)$-substrings. These observations indicate that care must be taken when including code-words which contain repeating s-substrings (indeed, where observations are made via the multiset of s'-substrings, for some $s' \leqslant s+1$). On the other hand, if every s-substring of x is unique, then x is uniquely reconstructible from the multiset of its $(s+1)$-substrings (and in fact, s'-substrings, for all $s' > s$), as evident from a greedy reconstruction algorithm (which at each stage searches for the next/previous character in the information string). This observation motivates the study of *repeat-free strings*; x is said to be s-repeat-free if every s-substring of x is unique (put differently, if x is of length n, then it contains $n-s+1$ distinct s-substrings).

Focus on repeat-free strings is further justified by the following results. It was observed in [17], via introduction of *profile vectors*, that over an alphabet of size q, where the length of strings n grows, if $s < \frac{\log_q(n)}{1+\epsilon}$ then the rate of all existing s-substring multisets vanishes. Conversely, it was demonstrated in [20] using probabilistic arguments that the asymptotic redundancy of the code-book consisting of all s-repeat-free strings of length n (which, as noted above, is an upper bound for the redundancy of a code assuring reconstruction from $(s+1)$-substrings), is $O(n^{2-s/\log_q(n)})$; thus, when $s > (1+\epsilon)\log_q(n)$, the rate of repeat-free strings alone is 1.

In this paper, we extend the setting of previous works by allowing information strings to suffer a bounded number of errors, prior to the sampling of their substrings. We study this model under two separate error models: substitution (Hamming) errors, and deletion errors. In both cases we show (see Theorems 7 and 13) that when $s > (1+\epsilon)\log_q(n)$ and the fraction of errors in the substring length s is sufficiently small, the rate of generalized repeat-free strings dubbed *resilient-repeat-free* suffers no penalty from the process of sampling, or from the presence of noise (when compared to the results of [20]); i.e., the required added redundancy is sub-linear. In the case of Hamming noise, we also show that when the fraction of errors is too large, resilient-repeat-free strings do not exist. However, it is left for future works to determine the precise transition between the two regimes.

It should be noted that [18] presented almost explicit en-/decoding algorithms for codes with a similar noise model. However, in that paper's setting, substitution noise affects individual substrings *after* sampling; the codes it constructs are capable of correcting a constant number of errors in each substring, but requires the assumption that errors do not affect the same information symbol in a majority of the substrings that reflect it. Therefore, its setting is incompatible with the one considered herein, whereby each error occurring *before* sampling affects s consecutive substrings. [21] also developed codes with full rate, capable of correcting a fixed number of errors, occurring in substrings independently after sampling. It replaced the aforementioned restriction by a constraint on the number of total erroneous substrings, which is at most logarithmic in the information string's length. Hence, the total number of errors in its setting remains asymptotically smaller than the one incurred in the setting considered here.

III. Preliminaries

Let Σ^* be the set of finite strings over an alphabet Σ, which we assume to be a finite unital ring of size q (e.g., $\mathbb{Z}_q$). For $x = x(0)x(1)\cdots x(n-1) \in \Sigma^*$, we let $|x| = n$ denote the *length* of x. We note that indices in the sequel are numbered $0, 1, \ldots$. For $x, y \in \Sigma^*$, we let xy be their concatenation. For $I \subseteq \mathbb{N}$ and $x \in \Sigma^*$, we denote by x_I the restriction of x to indices in I (excluding any indices $|x| \leqslant i \in I$), ordered according to the naturally inherited order on I.

We let $|A|$ denote the size of a finite set A. For a code $C \subseteq \Sigma^n$, we define its *redundancy* $\mathrm{red}(C) \triangleq n - \log_q|C|$, and *rate* $R(C) \triangleq \frac{1}{n}\log_q|C| = 1 - \frac{\mathrm{red}(C)}{n}$.

For $n \in \mathbb{N}$, denote $[n] \triangleq \{0, 1, \ldots, n-1\}$. Although perhaps confusable, for $m \leqslant n \in \mathbb{N}$ we use the common notation $[m, n] \triangleq \{m, m+1, \ldots, n\}$. We shall interpret x_I as enumerated by $[|I|]$, i.e., $x_I(0) = x(\min I)$, etc. Where it is convenient, we will also assume $I \subseteq \mathbb{N}$ to be enumerated by $[|I|]$, such that the order of elements is preserved; i.e., $I = \{I(i) : i \in [|I|]\}$, and for all $i \in [|I|-1]$ one has $I(i) < I(i+1)$. Under this convention, e.g., $x_I(0) = x(I(0))$. We follow the standard ring notation in denoting for $j \in \mathbb{N}$ and $I \subseteq \mathbb{N}$ the *coset* $j + I \triangleq \{j + i : i \in I\}$.

For $x \in \Sigma^*$ and $i, s \in \mathbb{N}$, where $i + s \leqslant |x|$, we say that $x_{i+[s]}$ is the length s *substring* of x at index i, or *s-mer* (at index i), or *s-gram*, for short. Using notation from [14], for $x \in \Sigma^*$ and $s \in \mathbb{N}$ we denote the multiset of s-mers of x by

$$Z_s(x) \triangleq \left\{\left\{x_{i+[s]} : 0 \leqslant i \leqslant |x| - s\right\}\right\}.$$

We follow [20] in denoting the set of *s-repeat-free* strings

$$\mathcal{RF}_s(n) \triangleq \left\{x \in \Sigma^n : i < j \implies x_{i+[s]} \neq x_{j+[s]}\right\}.$$

Assuming an underlying error model, known in context but yet to be determined, we let $S^t(x)$, for some $x \in \Sigma^*$, be the set of strings $y \in \Sigma^*$ that may be the product of t errors occurring

to x. We let $B^t(x) \triangleq \bigcup_{t' \leqslant t} S^{t'}(x)$. Using this notation, our aim shall be to study and design codes $C \subseteq \Sigma^n$, such that given $x \in C$ and $y \in B^t(x)$, for some fixed (or bounded) t, x can be uniquely reconstructed given only $Z_s(y)$. We shall study constraints which allow unique reconstruction of y, and state in Corollary 9 specific cases where this in turn allows reconstruction of x.

IV. Resilient-repeat-free strings

A. Substitution noise

In this section we consider substitution noise, with error spheres $S^t_{\mathrm{s}}(x) \triangleq \{y : d_{\mathrm{H}}(y,x) = t\}$, where $d_{\mathrm{H}}(x,y)$ denotes the Hamming distance between x and y.

We present and study a family of repeat-free strings which are resilient to substitution errors:

Definition 1 *We say that $x \in \Sigma^*$ is (t,s)-*resilient repeat free *if the result of any t substitution errors to x is s-repeat-free. More precisely, we define*

$$\mathcal{RRF}^{\mathrm{s}}_{t,s}(n) \triangleq \big\{x \in \Sigma^n : B^t_{\mathrm{s}}(x) \subseteq \mathcal{RF}_s(n)\big\}.$$

Lemma 2 *Take $t \in \mathbb{N}$, $x \in \Sigma^n$. If for all $0 \leqslant i < j \leqslant n - s$*

$$d_{\mathrm{H}}\big(x_{i+[s]}, x_{j+[s]}\big) > t + \max\{0, \min\{t, s-j+i\}\},$$

then $x \in \mathcal{RRF}^{\mathrm{s}}_{t,s}(n)$.

The proof follows from applying the triangle inequality by cases on $(i+[s]) \cap (j+[s])$, and is delegated to the full version of the manuscript.

Definition 3 *For positive $s \leqslant n$, denote $\binom{[n]}{s} \subseteq 2^{[n]}$ the collection of s-subsets of $[n]$. A pair of subsets $(I,J) \in \binom{[n]}{s}^2$ is said to be* observable *if $I(k) < J(k)$ for all $k \in [s]$.*

Given a string $x \in \Sigma^n$, known from context, we will denote for an observable pair $(I,J) \in \binom{[n]}{s}^2$

$$u_{I,J} \triangleq x_I - x_J \in \Sigma^s.$$

We also denote $L_I \triangleq \{(P,Q) : (P,Q)$ is observable, $(P \cup Q) \cap I = \emptyset\}$. To simplify notation, where some $s \leqslant n$ is also given, we shall abbreviate $u_{i,j} \triangleq u_{i+[s],j+[s]}$ and $L_i \triangleq L_{i+[s]}$, for any $0 \leqslant i < j \leqslant n - s$.

Lemma 4 *Take $s \leqslant n$ and an observable pair $(I,J) \in \binom{[n]}{s}^2$. Further, let $x \in \Sigma^n$ be chosen uniformly at random. Then $u_{I,J}$ is distributed uniformly and mutually independent of $\{u_{P,Q} : (P,Q) \in L_I\}$.*

Proof: First, since $u_{I,J}$ is the image of x under a linear map (more precisely, a module homomorphism), the pre-image of any point is a coset of map's kernel and, thus, of equal size; as a result, $u_{I,J}$ is distributed uniformly on the map's range. Since (I,J) is observable, the map is surjective onto Σ^s, hence the first part is completed.

Second, observe that x_I is independent of $x_{[n]\setminus I}$, hence mutually independent of $\{u_{P,Q} : (P,Q) \in L_I\}$. Since given $x_{[n]\setminus I}$, there exist a bijection between x_I and $u_{I,J}$, the proof is concluded. ∎

Before stating Lemma 6, on which the main results of this section rely, we make a few additional notations. First, fix $a > 1$, and denote $s_a = \lfloor a \log_q(n) \rfloor$ as n grows. By slight abuse of notation, we let $\mathcal{RRF}^{\mathrm{s}}_{t,a}(n) \triangleq \mathcal{RRF}^{\mathrm{s}}_{t,s_a}(n)$, and $\mathcal{RRF}^{\mathrm{s}}_{t,a} \triangleq \bigcup_{n\in\mathbb{N}} \mathcal{RRF}^{\mathrm{s}}_{t,a}(n)$. Second, for $0 < k \leqslant s_a$, denote

$$\mathcal{A}_k \triangleq \big\{x \in \Sigma^{s_a+k} : \exists y \in B^t_{\mathrm{s}}(x) : y_{[s_a]} = y_{k+[s_a]}\big\}.$$

Finally, denote $\pi_k \triangleq \Pr(x \in \mathcal{A}_k)$, where $x \in \Sigma^{s_a+k}$ is chosen uniformly at random.

Corollary 5 $\pi_k \leqslant q \cdot n^{-a\left(1-H_q\left(\delta+\min\left\{\delta,1-\frac{k}{s_a}\right\}\right)\right)}$.

Proof: By Lemma 4 $u_{i,j} \in \Sigma^{s_a}$ is distributed uniformly, hence for $w < \frac{q-1}{q}s_a$ it follows from, e.g., [30, Lem. 4.7]), that $\Pr(\mathrm{wt}(u_{i,j}) \leqslant w) \leqslant q^{s_a(H_q(w/s_a)-1)} \leqslant q \cdot n^{-a(1-H_q(w/s_a))}$. The claim now follows from Lemma 2. ∎

Lemma 6 *Denote $\pi' \triangleq \max_{0<k<s_a} \pi_k$. If $2s_a n \pi_{s_a}, 4s_a^2\pi' < 1/2e$ then, as $n \to \infty$,*

$$\mathrm{red}\big(\mathcal{RRF}^{\mathrm{s}}_{t,a}(n)\big) = O\big(n\log(n)\pi' + n^2\pi_{s_a}\big).$$

Proof: We define for all $0 \leqslant i < j \leqslant n - s_a$ the sets

$$A_{i,j} \triangleq \big\{x \in \Sigma^n : \exists y \in B^t_{\mathrm{s}}(x) : y_{i+[s_a]} = y_{j+[s_a]}\big\}.$$

Note that $\Sigma^n \setminus \mathcal{RRF}^{\mathrm{s}}_{t,a}(n) = \bigcup_{i,j} A_{i,j}$.

We let $x \in \Sigma^n$ be chosen uniformly at random. Then $\Pr(x \in A_{i,j}) = \pi_{\min\{s_a, j-i\}}$. Further,

$$\big|\mathcal{RRF}^{\mathrm{s}}_{t,a}(n)\big| = q^n \cdot \Pr\big(x \in \mathcal{RRF}^{\mathrm{s}}_{t,a}(n)\big),$$

and hence

$$\mathrm{red}\big(\mathcal{RRF}^{\mathrm{s}}_{t,a}(n)\big) = -\log_q \Pr\big(x \in \mathcal{RRF}^{\mathrm{s}}_{t,a}(n)\big).$$

Our proof strategy relies on Lovász's local lemma (LLL) [29], as follows. If for all $0 \leqslant i < j \leqslant n - s_a$ there exist constants $0 < f_{i,j} < 1$ such that

$$\Pr(x \in A_{i,j}) \leqslant f_{i,j} \prod_{\Gamma} (1 - f_{p,q}),$$

where Γ is such that $\{x \in A_{i,j}\}$ is mutually independent of $\{\{x \in A_{p,q}\} : (p,q) \notin \Gamma\}$, then the lemma states that

$$\Pr\Big(x \notin \bigcup\nolimits_{i,j} A_{i,j}\Big) \geqslant \prod_{i,j} (1 - f_{i,j}).$$

Note that, in our notation, $\Pr\Big(x \notin \bigcup_{i,j} A_{i,j}\Big) = \Pr\big(x \in \mathcal{RRF}^{\mathrm{s}}_{t,a}(n)\big)$ and $\Pr(x \in A_{i,j}) = \pi_{\min\{s_a, j-i\}}$.

To determine Γ, we claim for $0 \leqslant i < j \leqslant n - s_a$ that the event $\{x \in A_{i,j}\}$ is mutually independent of the events $\{\{x \in A_{p,q}\} : |i-p|, |i-q| \geqslant s_a\}$. Indeed, Lemma 4 then implies that $u_{i,j}$ is mutually independent of $\{u_{p,q} : (p,q) \in L_i\}$. It is left to the reader to verify that there exists a set

$B_{i,j} \subseteq \Sigma^{s_a}$, which depends on i,j but not x, such that the event $\{x \in A_{i,j}\}$ can be restated as $\{u_{i,j} \in B_{i,j}\}$ (and similarly for p,q); therefore, the claim holds. Hence, for any $0 \leqslant i < j \leqslant n - s_a$ we let $\Gamma \triangleq \{(p,q) : (p,q) \notin L_i\}$, and observe that it consists of strictly less than $4s_a^2$ pairs (p,q) such that $|p-q| < s_a$, and strictly less than $2s_a n$ other pairs.

Recalling for $0 < f < 1$ that $1 - f \geqslant e^{-f/(1-f)}$, and denoting $y = f/(1-f)$ we observe

$$\begin{aligned} f_{i,j}\prod_{p,q}(1-f_{p,q}) &= \tfrac{f_{i,j}}{1-f_{i,j}} \prod_{p,q \text{ inc. } i,j}(1-f_{p,q}) \\ &\geqslant y_{i,j}\exp\Bigg(-\sum_{p,q \text{ inc. } i,j} y_{p,q}\Bigg) \end{aligned}$$

We shall apply an almost symmetric version of LLL, where

$$f_{i,j} = \begin{cases} f, & j - i \geqslant s_a, \\ f', & j - i < s_a. \end{cases}$$

Then, to satisfy the LLL conditions it suffices that $\pi_{s_a} \leqslant ye^{-4s_a^2 y' - 2s_a n y}$ and $\pi' \leqslant y' e^{-4s_a^2 y' - 2s_a n y}$, where y, y' are defined as above for f, f' respectively. Let $y' \triangleq e\pi'$, $y \triangleq e\pi$. We have,

$$\begin{aligned} ye^{-4s_a^2 y' - 2s_a n y} &= \pi e^{1 - e\left(4s_a^2\pi' + 2s_a n \pi_{s_a}\right)} > \pi; \\ y'e^{-4s_a^2 y' - 2s_a n y} &= \pi' e^{1 - e\left(4s_a^2\pi' + 2s_a n \pi_{s_a}\right)} > \pi', \end{aligned}$$

as required.

Finally,

$$\begin{aligned} \operatorname{red}\big(\mathcal{RRF}^{\mathrm{s}}_{t,a}(n)\big) &= -\log_q \Pr\big(x \in \mathcal{RRF}^{\mathrm{s}}_{t,a}(n)\big) \\ &\leqslant -\log_q \Pr\Big(x \notin \textstyle\bigcup_{i,j} A_{i,j}\Big) \\ &\leqslant -\log_q \textstyle\prod_{i,j}(1-f_{i,j}) \\ &= \textstyle\sum_{i,j} \log_q(1+y_{i,j}) \\ &\leqslant \tfrac{1}{\ln(q)} \textstyle\sum_{i,j} y_{i,j} \\ &\leqslant \tfrac{ns_a}{\ln(q)} y' + \tfrac{n^2}{\ln(q)} y, \end{aligned}$$

which concludes the proof. ∎

Before utilizing Lemma 6 to find the redundancy of resilient-repeat-free strings, we make the following additional notation. For a fixed real number $\delta > 0$, we denote $t_\delta \triangleq \lfloor \delta s_a \rfloor = \big\lfloor \delta \big\lfloor a \log_q(n) \big\rfloor \big\rfloor$; as before, we abbreviate $\mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n) \triangleq \mathcal{RRF}^{\mathrm{s}}_{t_\delta, s_a}(n)$ (and similarly $\mathcal{RRF}^{\mathrm{s}}_{\delta,a}$).

Theorem 7 *Fix $a > 1$, $0 < \delta < \frac{q-1}{2q}$. Then, as $n \to \infty$,*

$$\operatorname{red}\big(\mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n)\big) = O\big(n^{2 - a(1 - H_q(\delta))}\big).$$

Proof: If $a \leqslant (1 - H_q(\delta))^{-1}$ the proposition vacuously holds.

Otherwise, let $x \in \Sigma^{s_a + k}$ be chosen uniformly at random. Based on Corollary 5, we observe for $\delta < \frac{q-1}{2q}$ and sufficiently large n that $2s_a n \pi_{s_a}, 4s_a^2 \max_{0<k<s_a}(\pi_k) < 1/2e$, satisfying the conditions of Lemma 6.

Since for all $k < s_a$ we also have $n \log n \pi_k = O\big(n^{2 - a(1 - H_q(\delta))}\big)$, the claim follows from Lemma 6. ∎

Corollary 8 *Take $0 < \delta < \frac{q-1}{2q}$. If $a > (1 - H_q(\delta))^{-1}$ then $R\big(\mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n)\big) = 1 - o(1)$, and if $a \geqslant 2(1 - H_q(\delta))^{-1}$, then $\mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n)$ incurs a constant number of redundant symbols.*

The last corollary can be viewed in the context of related works; as mentioned above, [20] demonstrated that, letting $s = \lfloor a \log_q n \rfloor$, if $a > 1$ then $R(\mathcal{RF}_s(n)) = 1 - o(1)$, and if $a \geqslant 2$ then $\operatorname{red}(\mathcal{RF}_s(n)) = O(1)$. Corollary 8 demonstrates that if $a > 1$ (respectively $a > 2$), then for all sufficiently small $\delta > 0$ it holds that $R\big(\mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n)\big) = 1 - o(1)$ (respectively, $\operatorname{red}\big(\mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n)\big) = O(1)$). I.e., resilient-repeat-free sequences for a number of substitutions errors logarithmic in the string length (linear in the substring length) incur no additional asymptotic cost.

Based on the last corollary, we can now demonstrate the existence of error-correcting codes for the noisy substring channel, which achieve at most a constant redundancy over that of classical error-correcting codes for Hamming noise.

Corollary 9 *Let $C \subseteq \Sigma^n$ be an error-correcting code, capable of correcting t_δ substitution errors, and denote, for some $z \in \Sigma^n$, $\bar{C}_z \triangleq (z + C) \cap \mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n)$. Then for any $x \in \bar{C}_z$ and $y \in B^{t_\delta}_{\mathrm{s}}(x)$, it is possible to uniquely decode x observing only $Z_{s_a+1}(y)$. Further, decoding is possible through a greedy algorithm for reconstruction of y, followed by application of any decoding scheme for C.*

Further, in the cases indicated in Corollary 8, where $\operatorname{red}\big(\mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n)\big) = O(1)$, there exists z satisfying $\operatorname{red}(\bar{C}_z) = \operatorname{red}(C) + O(1)$.

Note that Corollary 9 is unfortunately nonconstructive. We leave the interesting problem of constructing an encoder for error-correcting codes in $\mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n)$ (and in particular an efficient encoder) for future work.

Definition 10 *For a real δ, $0 \leqslant \delta < 1$, and an integer $s > 0$, let $M_q(s,\delta)$ be the maximum number of code-words in a code $C \subseteq \Sigma^s$ such that $d_{\mathrm{H}}(x,y) \geqslant \delta s$ for any distinct $x, y \in C$. For a given $\delta > 0$, define the maximum achievable rate by*

$$R_q(\delta) \triangleq \limsup_{k \to \infty} \tfrac{1}{s} \log_q M_q(s,\delta).$$

For completeness, we state the well-known GV and EB bounds (see, e.g., [30, Thm.4.9-12]) for $\delta \leqslant \frac{q-1}{q}$,

$$1 - H_q(\delta) \leqslant R_q(\delta) \leqslant 1 - H_q\Big(\tfrac{q-1}{q}\Big(1 - \sqrt{1 - \tfrac{q}{q-1}\delta}\Big)\Big).$$

The following lemma states a converse bound to Corollary 8.

Lemma 11 *If $t \geqslant \delta s_a$ and $a < R_q(\delta)^{-1}$, then for sufficiently large $n \in \mathbb{N}$*

$$\mathcal{RRF}^{\mathrm{s}}_{t,a}(n) = \emptyset.$$

In particular, the statement holds if $t \geqslant \frac{q-1}{q} s_a$, for all a.

Proof: Take, on the contrary, some $x \in \mathcal{RRF}^{\mathrm{s}}_{t,a}(n)$. By Definition 1, the s_a-mers

$$\left\{x_{is_a+[s_a]} : 0 \leqslant i \leqslant \lfloor n/s_a \rfloor - 1\right\} \subseteq \Sigma^{s_a}$$

form a code of size $\lfloor n/s_a \rfloor$ and minimum distance $d > t \geqslant \delta s_a$. By Definition 10, for $n \to \infty$ we obtain

$$\frac{\log_q \lfloor n/s_a \rfloor}{s_a} \leqslant R_q(\delta) + o(1).$$

Recalling $s_a = \left\lfloor a \log_q n \right\rfloor$ yields that

$$\tfrac{1}{a} \leqslant R_q(\delta) + o(1).$$

For sufficiently large n, the latter violates the condition imposed in the statement. ∎

It should be noted that Lemma 11 specifically pertains to resilient-repeat-free strings, which the reader will observe are not strictly required for successful reconstruction of information (although in the noiseless case, they achieve optimum asymptotic rate).

Before concluding, we note that a twofold gap remains between Theorem 7 and the converse of Lemma 11. First, $\mathrm{red}\big(\mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n)\big)$ is not characterized when $R_q(\delta)^{-1} \leqslant a \leqslant (1 - H_q(\delta))^{-1}$; and second, it is not found when $\delta > \frac{q-1}{2q}$. In the full version of the manuscript, we show that Lemma 6 may be used to prove $\mathrm{red}\big(\mathcal{RRF}^{\mathrm{s}}_{\delta,a}(n)\big) = o(n) + O(n^{2-a(1-H_q(\delta))})$ for all $\delta < \delta^*_q$, where $\frac{q-1}{2q} < \delta^*_q \leqslant \frac{q-1}{q}$.

B. Deletion noise

In this section, we consider deletion noise instead of substitution errors. For $x \in \Sigma^n$, let $S^t_{\mathrm{d}}(x) \subseteq \Sigma^{n-t}$ denote the set of strings generated from x by t deletions.

Definition 12 *As in the previous section, we define a family of repeat-free strings which is resistant to deletion noise:*

$$\mathcal{RRF}^{\mathrm{d}}_{t,s}(n) \triangleq \left\{x \in \Sigma^n : S^t_{\mathrm{d}}(x) \subseteq \mathcal{RF}_s(n-t)\right\}.$$

We denote $s_a \triangleq \lfloor a \log_q(n) \rfloor$ and $t_\delta \triangleq \lfloor \delta s_a \rfloor$, for some fixed real numbers $a > 1$ and $\delta > 0$. As in the previous sections, we denote $\mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n) \triangleq \mathcal{RRF}^{\mathrm{d}}_{t_\delta,s_a}(n)$ and $\mathcal{RRF}^{\mathrm{d}}_{\delta,a} \triangleq \bigcup_{n\in\mathbb{N}} \mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n)$. Then we have the following:

Theorem 13 *For all $a > 1$ and $\delta > 0$ it holds that*

$$\mathrm{red}\Big(\mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n)\Big) = O\Big(n^{2-a+\frac{2a(1+\delta)}{\log_2(q)} H_2(\delta/(1+\delta))}\Big).$$

Proof: We follow a similar strategy as in Lemma 6, but apply a symmetric bound. Note that a sufficient condition for $x \in \mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n)$ is that for every observable pair $(I,J) \in \binom{[n]}{s_a}^2$, such that $I(s_a - 1) - I(0) < s_a + t_\delta$ (and similarly for J), it holds that $x_I \neq x_J$. For such a pair, denote $A_{I,J} \triangleq \{x \in \Sigma^n : x_I = x_J\} = \{x \in \Sigma^n : u_{I,J} = 0\}$. Again, we let $x \in \Sigma^n$ be chosen uniformly at random, and use the Lovász local lemma to bound $\Pr\Big(x \in \mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n)\Big) \geqslant \Pr\Big(x \notin \bigcup_{I,J} A_{I,J}\Big)$ from below.

For any observable pair (I,J), note that $\Pr(x \in A_{I,J}) = q^{-s_a} \leqslant q \cdot n^{-a}$, and for convenience denote $\pi \triangleq q \cdot n^{-a}$.

Next, we estimate $|\Gamma|$ (a set of observable pairs (P,Q) satisfying the same properties) such that the event $\{x \in A_{I,J}\}$ is mutually independent of $\{\{x \in A_{P,Q}\} : (P,Q) \notin \Gamma\}$. Observe by Lemma 4 that it suffices that Γ consists of all $(P,Q) \notin L_I$ with the above properties. Thus, to determine P, it suffices to choose 1) a single element of I (which shall be a member of $P \cap I$); 2) an interval of length $s_a + t_\delta$ containing the chosen element; 3) any $s_a - 1 < s_a$ additional elements of the chosen interval. Then Q can be chosen from any interval of length $s_a + t_\delta$. The same holds for a suitable choice of $Q \cap I \neq \emptyset$. Thus, $|\Gamma| \leqslant s_a(s_a + t_\delta) n \binom{s_a+t_\delta}{s_a}^2$.

Now, to apply LLL, we find $0 < f < 1$ such that $\pi \leqslant f(1-f)^{|\Gamma|}$. From $\binom{k}{l} \leqslant 2^{kH_2(l/k)}$ (a relaxation of, e.g., [31, Ch.10, Sec.11, Lem.7]) we observe that $\binom{s_a+t_\delta}{t_\delta}^2 \leqslant n^{\frac{2a(1+\delta)}{\log_2(q)} H_2(\delta/(1+\delta))}$. If $\frac{2(1+\delta)}{\log_2(q)} H_2(\delta/(1+\delta)) \geqslant 1$ or $a \leqslant \Big(1 - \frac{2(1+\delta)}{\log_2(q)} H_2(\delta/(1+\delta))\Big)^{-1}$, the proposition vacuously holds. Otherwise, we note that $(|\Gamma|+1)\pi = o(1)$ as $n \to \infty$. Let $y \triangleq e\pi$, and for sufficiently large n note $ye^{-(|\Gamma|+1)y} = \pi e^{1-e(|\Gamma|+1)\pi} > \pi$. By further denoting $f \triangleq \frac{y}{1+y}$, and recalling for $0 < f < 1$ that $1 - f \geqslant e^{-f/(1-f)}$, we observe

$$f(1-f)^{|\Gamma|} \geqslant \tfrac{f}{1-f} \exp\Big(-\tfrac{f}{1-f}(|\Gamma|+1)\Big) = ye^{-(|\Gamma|+1)y} > \pi.$$

Finally, one needs also note that the number of observable pairs (I,J) satisfying the given requirements is no more than $\binom{n-s_a-t_\delta}{2} \cdot \binom{s_a+t_\delta}{s_a}^2 \leqslant \frac{1}{2} n^2 \binom{s_a+t_\delta}{s_a}^2$. By LLL it follows that

$$\Pr\left(x \notin \bigcup_{I,J} A_{I,J}\right) \geqslant (1-f)^{\frac{n^2}{2}\binom{s_a+t_\delta}{s_a}^2} = (1+y)^{-\frac{n^2}{2}\binom{s_a+t_\delta}{s_a}^2},$$

and hence

$$\begin{aligned}
\mathrm{red}\Big(\mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n)\Big) &= -\log_q \Pr\Big(x \in \mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n)\Big) \\
&\leqslant \frac{n^2}{2}\binom{s_a+t_\delta}{s_a}^2 \log_q(1+y) \\
&\leqslant \frac{1}{2\ln(q)}\binom{s_a+t_\delta}{s_a}^2 n^2 y \\
&= \frac{q}{2\ln(q)}\binom{s_a+t_\delta}{s_a}^2 n^{2-a},
\end{aligned}$$

which completes the proof. ∎

Corollary 14 *If $a > \Big(1 - \frac{2(1+\delta)}{\log_2(q)} H_2(\delta/(1+\delta))\Big)^{-1}$, for any $\delta > 0$, then $R(\mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n)) = 1 - o(1)$, and if $a \geqslant 2\Big(1 - \frac{2(1+\delta)}{\log_2(q)} H_2(\delta/(1+\delta))\Big)^{-1}$ then $\mathrm{red}\Big(\mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n)\Big) = O(1)$.*

Note again that if $a > 1$ (respectively $a > 2$), then for all sufficiently small $\delta > 0$ it holds that $R\Big(\mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n)\Big) = 1 - o(1)$ (respectively, $\mathrm{red}\Big(\mathcal{RRF}^{\mathrm{d}}_{\delta,a}(n)\Big) = O(1)$). Before concluding, we also note that a parallel statement to Corollary 9 holds in this setting, as well.

REFERENCES

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

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

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

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

[10] 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.

[11] J. Sima, N. Raviv, and J. Bruck, "On coding over sliced information," in *Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT'2019), Paris, France*, Jul. 2019, pp. 767–771.

[12] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage," *IEEE Trans. on Inform. Theory*, vol. 66, no. 4, pp. 7682–7696, Apr. 2020.

[13] J. Sima, N. Raviv, and J. Bruck, "Robust indexing - optimal codes for DNA storage," in *Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT'2020), Los Angeles, CA, USA*, Jun. 2020, pp. 717–722.

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

[15] 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.

[16] 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.

[17] Z. Chang, J. Chrisnata, M. F. Ezerman, and H. M. Kiah, "Rates of DNA sequence profiles for practical values of read lengths," *IEEE Trans. on Inform. Theory*, vol. 63, no. 11, pp. 7166–7177, Nov. 2017.

[18] R. Gabrys and O. Milenkovic, "Unique reconstruction of coded strings from multiset substring spectra," *IEEE Trans. on Inform. Theory*, vol. 65, no. 12, pp. 7682–7696, Dec. 2019.

[19] J. Chrisnata, H. M. Kiah, S. Rao, A. Vardy, E. Yaakobi, and H. Yao, "On the number of distinct k-decks: Enumeration and bounds," in *Proceedings of the 2019 19th International Symposium on Communications and Information Technologies (ISCIT), Ho Chi Minh City, Vietnam, Vietnam*, Sep. 2019, pp. 519–524.

[20] O. Elishco, R. Gabrys, M. Médard, and E. Yaakobi, "Repeat-free codes," in *Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT'2019), Paris, France*, Jul. 2019, pp. 932–936.

[21] S. Marcovich and E. Yaakobi, "Reconstruction of strings from their substrings spectrum," in *Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT'2020), Los Angeles, CA, USA*, Jun. 2020, pp. 658–663.

[22] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Anchor-based correction of substitutions in indexed sets," in *Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT'2019), Paris, France*, Jul. 2019, pp. 757–761.

[23] 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.

[24] N. Beeri and M. Schwartz, "Improved rank-modulation codes for dna storage with shotgun sequencing," *arXiv preprint arXiv:2101.06033*, 2021.

[25] S. Jain, F. Farnoud, M. Schwartz, and J. Bruck, "Noise and uncertainty in string-duplication systems," in *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT'2017), Aachen, Germany*, Jun. 2017, pp. 3120–3124.

[26] 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.

[27] F. Farnoud, M. Schwartz, and J. Bruck, "Estimation of duplication history under a stochastic model for tandem repeats," *BMC Bioinformatics*, vol. 20, no. 1, pp. 64–74, Feb. 2019.

[28] Y. Yehezkeally and N. Polyanskii, "On codes for the noisy substring channel," *arXiv preprint arXiv:2102.01412*, 2021.

[29] J. Spencer, "Asymptotic lower bounds for Ramsey functions," *Discrete Mathematics*, vol. 20, pp. 69–76, 1977. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0012365X77900449

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

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