Bounds On the Maximum Cardinality of Indel and Substitution Correcting Codes

Spee, Ward J. P.; Weber, Jos H.

DOI
10.1109/TMBMC.2024.3388971

Publication date
2024

Document Version
Final published version

Published in
IEEE Transactions on Molecular, Biological, and Multi-Scale Communications

Bounds on the Maximum Cardinality of Indel and Substitution Correcting Codes

Ward J. P. Spee, *Student Member, IEEE*, and Jos H. Weber, *Senior Member, IEEE*

***Abstract*—Recent advances in DNA data storage have attracted renewed attention towards deletion, insertion and substitution correcting codes. Compared to codes aimed at correcting either substitution errors or deletion and insertion (indel) errors, the understanding of codes that correct combinations of substitution and indel errors lags behind. In this paper, we focus on the maximal size of q-ary t-indel s-substitution correcting codes.Our main contributions include two Gilbert-Varshamov inspired lower bounds on this size. On the upper bound side, we prove a Singleton-like bound, a family of sphere-packing upper bounds and an integer linear programming bound. Several of these bounds are shown to improve upon existing results. Moreover, we use these bounds to derive a lower bound and an upper bound on the asymptotic redundancy of maximally sized t-indel s-substitution correcting codes.**

***Index Terms*—DNA data storage, error correcting codes, indels, deletions, insertions, substitutions, Gilbert-Varshamov bound, Singleton bound, sphere-packing bound, ILP bound.**

I. Introduction

ADVANCES in practical research on DNA data storage in the last decade have shown that it could become a viable alternative for traditional archival storage methods in the future [1], [2], [3], [4], [5]. DNA molecules consisting of strings of nucleotides Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) are suited for this purpose because of their high-information density and longevity properties [6]. Error analysis of current DNA storage systems has revealed that the strings incur deletion, insertion and substitution errors during synthesis (writing), storage and sequencing (reading) [7]. Therefore, coding techniques that reduce error rates and correct combinations of these errors are essential for reliable use of DNA data storage systems.

Error rate reduction and error correction can be achieved by imposing restrictions on the set of quaternary strings that is used to convey information. For example, strings should not contain long homopolymer runs or a strongly unbalanced number of C/G nucleotides in comparison to A/T nucleotides, since these properties significantly increase error rates [7]. Although these restrictions are beneficial to reduce error rates, adherence does not eliminate the need for error correction. Error correction capabilities can be enforced by carefully selecting a subset of quaternary strings, i.e., a code that satisfies certain distance requirements. Henceforth, we will consider codes as a subset of q-ary strings for some alphabet consisting of $q \geq 2$ symbols, and treat coding for DNA data storage as a special case, i.e., $q = 4$.

For optimizing the efficiency of codes, it is interesting to study the maximum size of a code that can correct combinations of deletion, insertion and substitution errors. Classical error correcting codes aimed only at correcting substitution errors have been well-studied for over 75 years [8]. This has led to numerous non-asymptotic bounds on the maximum size of substitution correcting codes (see, e.g., [9], [10], [11], [12], [13]). The study of deletion and insertion (indel) correcting codes was initiated by Levenshtein [14]. He showed that a code that is able to correct t deletions (or insertions) is able to correct any t' deletions and t'' insertions, whenever $t'+t'' \leq t$. In other words, a t-deletion (insertion) correcting code is also a t-indel correcting code. This property shows the equivalence between correcting deletions and insertions, which warrants the terminology of t-indel correcting codes. Non-asymptotic bounds on the maximum size of t-indel correcting codes have been provided in e.g., [14], [15], [16], [17], [18].

In comparison with either substitution correcting codes or indel correcting codes, non-asymptotic bounds on the maximal cardinality of t-indel s-substitution correcting codes have been studied to a lesser degree in literature. Smagloy et al. [19] derived two upper bounds on this size for specific values of t and s. Several t-indel s-substitution correcting codes have been constructed, e.g., in [20], [21], which naturally imply non-asymptotic lower bounds. Moreover, note that each $(t + 2s)$-indel correcting codes is also a t-indel s-substitution correcting code, because a substitution can be seen as a deletion followed by an insertion. Hence, lower bounds on the maximum size of $(t + 2s)$-indel correcting codes imply lower bounds for t-indel s-substitution correcting codes as well.

The last observation that any $(t + 2s)$-indel correcting code is also a t-indel s-substitution correcting code might raise the preliminary question whether it is superfluous to consider the correction of substitutions separately. However, there are two arguments in favor of separating indel correction from substitution correction. First, it was recognized by Song et al. [20] that $(t + 2s)$-indel correcting codes are not necessarily optimal within the set of t-indel s-substitution correcting codes in

Manuscript received 22 September 2023; revised 26 January 2024; accepted 4 April 2024. Date of publication 16 April 2024; date of current version 17 June 2024. The associate editor coordinating the review of this article and approving it for publication was P. Siegel. *(Corresponding author: Ward J. P. Spee.)*

The authors are with the Department of Applied Mathematics, Delft University of Technology, 2600 AA Delft, The Netherlands (e-mail: ward.spee@hotmail.com; j.h.weber@tudelft.nl).

Digital Object Identifier 10.1109/TMBMC.2024.3388971

terms of redundancy.[1] Secondly, the error rates of indels and substitutions in DNA data storage are not necessarily equal [22]. Therefore, it is sensible to bound the number indels and substitutions by different parameters.

In this paper, we are interested in the maximum size of q-ary codes that correct combinations of indels and substitutions. The main contributions of this paper are several explicit and non-asymptotic bounds on this size that hold for a general number of indels and substitutions. More specifically, we present two lower bounds which are both inspired by the well-known Gilbert-Varshamov bound. Furthermore, we construct a Singleton-like upper bound, sphere-packing upper bounds and integer linear programming bounds.

The organisation of this paper is as follows. In Section II, notation, terminology and several prior results are discussed. Next, the aforementioned lower and upper bounds are derived in Sections III and IV, respectively. Lastly, two bounds are used in Section V in order to derive a lower and upper bound on the asymptotic redundancy of a maximally sized t-indel s-substitution correcting code.

II. Definitions and Preliminaries

For a finite set S, denote the cardinality of S by $|S|$. By convention, we set $\binom{a}{0} = 1$ for all $a \in \mathbb{Z}$, and $\binom{a}{b} = 0$ if either $a < 0$ and $b > 0$, or $b > a > 0$.

For convenience, we will view the four types of nucleotides as numerals using the following bijection,

$$\mathtt{A} \leftrightarrow 0,\ \mathtt{C} \leftrightarrow 1,\ \mathtt{T} \leftrightarrow 2,\ \mathtt{G} \leftrightarrow 3.$$

Hence, a single string of DNA of length n is represented by a word in the set $\{0,1,2,3\}^n$. More generally, we consider the alphabet with $q \geq 2$ symbols given by $\mathcal{B}_q \coloneqq \{0,1,\ldots,q-1\}$. The set of q-ary words (i.e., vectors) of length n with symbols from $\mathcal{B}_q$ is denoted by $\mathcal{B}_q(n) \coloneqq \{0,1,\ldots,q-1\}^n$. Let $n_1, n_2 \geq 1$ be integers, then we denote the concatenation of two words $\mathbf{u} \in \mathcal{B}_q(n_1)$ and $\mathbf{v} \in \mathcal{B}_q(n_2)$ by $(\mathbf{u}|\mathbf{v}) \in \mathcal{B}_q(n_1+n_2)$. A run in a word $\mathbf{x} \in \mathcal{B}_q(n)$ is a sequence of consecutive and identical symbols in $\mathbf{x}$ that is not contained within a longer such sequence. The number of runs in $\mathbf{x}$ is denoted by $r(\mathbf{x})$. For instance, let $\mathbf{y} = 11000322 \in \mathcal{B}_4(8)$, then $\mathbf{y}$ contains the runs 11, 000, 3 and 22, and hence $r(\mathbf{y}) = 4$. The number of words in $\mathcal{B}_q(n)$ with exactly r runs is given by [15],

$$|\{\mathbf{x} \in \mathcal{B}_q(n) : r(\mathbf{x}) = r\}| = q\binom{n-1}{r-1}(q-1)^{r-1}. \quad (1)$$

For integers $0 \leq t \leq n$ and $0 \leq s \leq n$, a code $\mathcal{C} \subseteq \mathcal{B}_q(n)$ is said to be a *t-indel s-substitution correcting code* if any q-ary word (not necessarily of length n) can be obtained from no more than one codeword by exactly t' deletions, t'' insertions and s or fewer substitutions, whenever $t' + t'' \leq t$. A 0-indel s-substitution correcting code is simply called an *s-substitution correcting code* and analogously a t-indel 0-substitution correcting code is called a *t-indel correcting code*. In order to maximize the amount of information that can be transmitted using a code, we are interested in the maximal size of a q-ary t-indel s-substitution correcting code with codewords of length n, which we denote by $M_q(n,t,s)$. The redundancy of a code $\mathcal{C}$ is defined by $n - \log_q(|\mathcal{C}|)$.

Denote by $\mathcal{V}_{t',t'',s}(\mathbf{x})$ the set of words that can be reached from $\mathbf{x} \in \mathcal{B}_q(n)$ by means of exactly t' deletions, t'' insertions and at most s substitutions. Clearly, the q-ary words in the set $\mathcal{V}_{t',t'',s}(\mathbf{x})$ have length $n - t' + t''$. Moreover, we define $\mathcal{D}_t(\mathbf{x}) = \mathcal{V}_{t,0,0}(\mathbf{x})$, $\mathcal{I}_t(\mathbf{x}) = \mathcal{V}_{0,t,0}(\mathbf{x})$ and $\mathcal{S}_s(\mathbf{x}) = \mathcal{V}_{0,0,s}(\mathbf{x})$. These sets often arise in the study of t-indel s-substitution correcting codes, because they allow for equivalent characterizations of these codes in terms of the set $\mathcal{V}_{t',t'',s}(\mathbf{x})$. The following lemma collects various equivalent characterizations from e.g., [20, Sec. II], [23, Lem. 2] and [19, Lem. 2].

Lemma 1: Let $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$ and $0 \leq s \leq n$ be integers, and let $\mathcal{C} \subseteq \mathcal{B}_q(n)$ be a code. Then, the following five statements are equivalent:

1) $\mathcal{C}$ is a t-indel s-substitution correcting code.
2) $\mathcal{V}_{t',t'',s}(\mathbf{c}_1) \cap \mathcal{V}_{t',t'',s}(\mathbf{c}_2) = \emptyset$ for all distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$, and for all integers $t', t'' \geq 0$ such that $t' + t'' \leq t$.
3) $\mathcal{V}_{t,0,s}(\mathbf{c}_1) \cap \mathcal{V}_{t,0,s}(\mathbf{c}_2) = \emptyset$ for all distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$.
4) $\mathcal{V}_{0,t,s}(\mathbf{c}_1) \cap \mathcal{V}_{0,t,s}(\mathbf{c}_2) = \emptyset$ for all distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$.
5) $\mathbf{c}_2 \notin \mathcal{V}_{t,t,2s}(\mathbf{c}_1)$ for all distinct $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$.

For general parameters t', t'' and s, and words $\mathbf{x} \in \mathcal{B}_q(n)$ determining the cardinality of $\mathcal{V}_{t',t'',s}(\mathbf{x})$ is a non-trivial task [24]. In the specific case that $t' = t'' = 0$ it holds that [8]

$$|\mathcal{S}_s(\mathbf{x})| = \sum_{i=0}^{s} \binom{n}{i}(q-1)^i, \quad (2)$$

for each $\mathbf{x} \in \mathcal{B}_q(n)$. The quantity $S^s_{n,q} \coloneqq \sum_{i=0}^{s}\binom{n}{i}(q-1)^i$ will be referred to as the size of the q-ary Hamming sphere of radius s. Moreover, it has been established [25] that

$$|\mathcal{I}_t(\mathbf{x})| = S^t_{n+t,q} = \sum_{i=0}^{t} \binom{n+t}{i}(q-1)^i. \quad (3)$$

Interestingly, the cardinalities of $\mathcal{S}_s(\mathbf{x})$ and $\mathcal{I}_t(\mathbf{x})$ depend on $\mathbf{x}$ only via the parameters n and q. In contrast, $|\mathcal{D}_t(\mathbf{x})|$ depends on the structure of the word $\mathbf{x}$ as well as the parameters n and q. To the best of authors' knowledge, an analytic formula of $|\mathcal{D}_t(\mathbf{x})|$ is not known for general t and therefore we have to rely on bounds. In [14], Levenshtein showed that

$$\binom{r(\mathbf{x})-t+1}{t} \leq |\mathcal{D}_t(\mathbf{x})| \leq \binom{r(\mathbf{x})+t-1}{t} \quad (4)$$

for all $\mathbf{x} \in \mathcal{B}_q(n)$. The lower bound was later improved by Hirschberg and Regnier [26] to

$$\sum_{i=0}^{t}\binom{r(\mathbf{x})-t}{i} \leq |\mathcal{D}_t(\mathbf{x})|. \quad (5)$$

An even stronger lower bound was found in [27], but it holds only for binary words. For $t \leq 5$, an analytic formula of $|\mathcal{D}_t(\mathbf{x})|$ has been provided in [28], but these expressions are rather involved for $t \geq 2$. Lastly, by using the observation that

[1] For instance, the single-substitution correcting binary Hamming code with words of length 7 has size 16 [11]. In contrast, in [17, Th. 1] it was shown that a binary two-indel correcting code has a maximal size of at most 11.

$\mathbf{x} \in \mathcal{I}_t(\mathbf{y})$ if and only if $\mathbf{y} \in \mathcal{D}_t(\mathbf{x})$, it was shown in [15] that the average cardinality of $\mathcal{D}_t(\mathbf{x})$ is given by

$$\frac{1}{q^n}\sum_{\mathbf{x}\in\mathcal{B}_q(n)}|\mathcal{D}_t(\mathbf{x})| = \frac{1}{q^n}\sum_{\mathbf{y}\in\mathcal{B}_q(n-t)}|\mathcal{I}_t(\mathbf{y})| \overset{(3)}{=} \frac{1}{q^t}\sum_{i=0}^{t}\binom{n}{i}(q-1)^i. \tag{6}$$

In the setting of combinations of indels and substitutions, we define the following quantity

$$V_{t,t,2s}^{avg} \coloneqq q^{-n}\sum_{\mathbf{x}\in\mathcal{B}_q(n)}\big|\mathcal{V}_{t,t,2s}(\mathbf{x})\big|. \tag{7}$$

Next, we establish the following simple lower bound on the cardinality of $\mathcal{V}_{t,0,s}(\mathbf{x})$.

Lemma 2: Let $n \geq 1$, $0 \leq t \leq \frac{n}{2}$ and $0 \leq s \leq \frac{n}{2}$ be integers. Let $\mathbf{x} \in \mathcal{B}_q(n)$ be a word with $r = r(\mathbf{x})$, then it holds that

$$\sum_{i=0}^{s}\binom{\lfloor\frac{n}{2}\rfloor}{i}(q-1)^i \cdot \sum_{j=0}^{t}\binom{\lfloor\frac{r}{2}\rfloor - t}{j} \leq \big|\mathcal{V}_{t,0,s}(\mathbf{x})\big|.$$

An analogous bound holds for the cardinality of $\mathcal{V}_{0,t,s}(\mathbf{x})$.

Lemma 3: Let $n \geq 1$, $0 \leq t \leq \frac{n}{2}$ and $0 \leq s \leq \frac{n}{2}$ be integers. Let $\mathbf{x} \in \mathcal{B}_q(n)$, then it holds that

$$\sum_{i=0}^{s}\binom{\lfloor\frac{n}{2}\rfloor}{i}(q-1)^i \cdot \sum_{i=0}^{t}\binom{\lceil\frac{n}{2}\rceil + t}{i}(q-1)^i \leq \big|\mathcal{V}_{0,t,s}(\mathbf{x})\big|.$$

We relegated the proofs of these lemmas to Appendices A & B. For $t = s = 1$, it is not necessary to use these bounds, since exact expressions are known. Namely, for integers $n \geq 1$ and $q \geq 2$, and any word $\mathbf{x} \in \mathcal{B}_q(n)$, it was stated in [19] that

$$\big|\mathcal{V}_{1,0,1}(\mathbf{x})\big| = \begin{cases} (n-1)(q-1)+1 & \text{if } r(\mathbf{x}) = 1, \\ r(\mathbf{x})\ (n-2)(q-1) & \\ \quad -r(\mathbf{x})+q+2 & \text{if } r(\mathbf{x}) \geq 2. \end{cases} \tag{8}$$

Exact expressions for the cardinalities of $\mathcal{V}_{0,1,1}(\mathbf{x})$ and $\mathcal{V}_{1,1,0}(\mathbf{x})$ were provided in [29] and [30], respectively. From previous expressions, it is apparent that the cardinality of $\mathcal{V}_{t',t'',s}(\mathbf{x})$ and $r(\mathbf{x})$ are strongly related. Hence, we provide a result that will be used repeatedly throughout this paper.

Lemma 4: Let $n \geq 1$, $q \geq 2$, $t' \geq 0$, $t'' \geq 0$ and $s \geq 0$ be integers such that $n - t' + t'' \geq 1$. Let $\mathbf{x} \in \mathcal{B}_q(n)$ and $\mathbf{y} \in \mathcal{V}_{t',t'',s}(\mathbf{x})$, then the following holds,

$$r(\mathbf{x}) - 2\big(t'+s\big) \ \leq r(\mathbf{y}) \leq r(\mathbf{x}) + 2\big(t''+s\big).$$

For a proof of this statement we refer to Appendix C.

Before proceeding to the main results of this paper, we make a remark about the way in which they will be presented. Unsurprisingly, several of the subsequent bounds on $M_q(n,t,s)$ are based on expressions or bounds on cardinality of $\mathcal{V}_{t',t'',s}(\mathbf{x})$. Since this cardinality is not known for all sets of parameters, to the best of the authors' knowledge, we first present implicit bounds that show the general dependency on $|\mathcal{V}_{t',t'',s}(\mathbf{x})|$. Next, these bounds are made explicit using existing results on $|\mathcal{V}_{t',t'',s}(\mathbf{x})|$. This two-part presentation has the advantage that the explicit bounds can be easily improved with the potential availability of new results on $|\mathcal{V}_{t',t'',s}(\mathbf{x})|$ in the future.

III. Gilbert-Varshamov Inspired Lower Bounds

The well-known Gilbert-Varshamov lower bound for s-substitution correcting codes [9], [10] is given by

$$M_q(n,0,s) \geq \frac{q^n}{\sum_{i=0}^{2s}\binom{n}{i}(q-1)^i}. \tag{9}$$

This bound is commonly proven using a sphere-covering argument where the spheres are given by $\mathcal{S}_{2s}(\mathbf{c})$ centered around the codewords $\mathbf{c} \in \mathcal{C}$ (see, e.g., [8, Th. 4.3]). In the case of substitutions, this proof is facilitated by the fact that these spheres are of equal size.

Tolhuizen [31] recognized that the Gilbert-Varshamov bound is also implied by Turán's theorem [32] from extremal graph theory. A particular consequence of the latter approach is that it easily generalizes to the case in which the spheres are not of equal size. For instance, this is the case for t-indel correcting codes when dealing with the spheres $\mathcal{V}_{t,t,0}(\mathbf{c})$. The approach from Tolhuizen was used by Levenshtein [15] to bound the maximal size of a t-indel correcting code from below. In particular, it was shown that

$$M_q(n,t,0) \geq \frac{q^{n+t}}{\Big(\sum_{i=0}^{t}\binom{n}{i}(q-1)^i\Big)^2}. \tag{10}$$

For completeness, we mention that other Gilbert-Varshamov related lower bounds on $M_q(n,t,0)$ are given in [33], [34]. In a different setting, multiple generalized Gilbert-Varshamov bounds are derived in [35] that resemble the following lemmas.

Next, it is a natural step to generalize the argument from Tolhuizen to t-indel s-substitution correcting codes.

Lemma 5: Let $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$ and $0 \leq s \leq n$ be integers. The following gives a lower bound on $M_q(n,t,s)$,

$$M_q(n,t,s) \geq \frac{q^n}{V_{t,t,2s}^{avg}}. \tag{11}$$

Proof: The idea of this proof is to translate the problem of finding a large code to the problem of finding a large clique.[2] This allows us to apply the argument from [31, Sec. II] to derive the desired lower bound on $M_q(n,t,s)$.

Define the undirected graph $G = (V, E)$ without loops or double edges as follows. Let $V = \mathcal{B}_q(n)$ be the set of nodes of G. Two distinct nodes $\mathbf{x}$ and $\mathbf{y}$ from V are joined by an edge in E if $\mathbf{x} \notin \mathcal{V}_{t,t,2s}(\mathbf{y})$. This is well-defined because it holds that $\mathbf{x} \notin \mathcal{V}_{t,t,2s}(\mathbf{y})$ if and only if $\mathbf{y} \notin \mathcal{V}_{t,t,2s}(\mathbf{x})$. Intuitively, the pairs of nodes that are connected by an edge can both be codewords in a t-indel s-substitution correcting code. The number of nodes equals $|V| = q^n$ and the number of edges is given by

$$\begin{aligned}|E| &= \frac{1}{2}\sum_{\mathbf{x}\in V}\big(|V| - |\mathcal{V}_{t,t,2s}(\mathbf{x})|\big) \\ &= \frac{1}{2}q^{2n} - \frac{1}{2}\sum_{\mathbf{x}\in\mathcal{B}_q(n)}|\mathcal{V}_{t,t,2s}(\mathbf{x})| = \frac{1}{2}q^n\Big(q^n - V_{t,t,2s}^{avg}\Big),\end{aligned}$$

where the first equality follows from the fact that each $\mathbf{x} \in V$ has $|V| - |\mathcal{V}_{t,t,2s}(\mathbf{x})|$ incident edges. Therefore, summing

[2] A clique of a graph G is an induced subgraph that is complete, i.e., all pairs of nodes are connected by an edge.

$|V| - |\mathcal{V}_{t,t,2s}(\mathbf{x})|$ over all nodes in $\mathbf{x} \in V$ equals $2|E|$ since each edge is counted twice. Observe that from the definition of the edges in G and Lemma 1 it follows that a clique of size k in G corresponds to a t-indel s-substitution correcting code $\mathcal{C}$ of size k.

Using the cardinalities of V and E it follows from the argument in [31, Sec. II] that there exists a clique in G of size $\lceil \frac{q^n}{V^{avg}_{t,t,2s}} \rceil$. For brevity, we do not repeat this argument here. In turn, this implies that there exists an equally large t-indel s-substitution correcting code, which concludes the proof. ∎

In essence, the strength of the lower bound on $M_q(n,t,s)$ from Lemma 5 is determined by the size of a clique that is implied by Turán's theorem. This observation allows us to improve Lemma 5 by using the stronger result of Caro [36] and Wei [37] on the size of the largest clique in a graph.

Lemma 6: Let $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$ and $0 \leq s \leq n$ be integers. The following gives a lower bound on $M_q(n,t,s)$,

$$M_q(n,t,s) \geq \sum_{\mathbf{x} \in \mathcal{B}_q(n)} \frac{1}{|\mathcal{V}_{t,t,2s}(\mathbf{x})|}. \tag{12}$$

Proof: Let $G = (V, E)$ be the graph as defined in the proof of Lemma 5. The degree of a node $\mathbf{x} \in V = \mathcal{B}_q(n)$ is given by

$$deg(\mathbf{x}) = \big|\mathcal{B}_q(n) \setminus \mathcal{V}_{t,t,2s}(\mathbf{x})\big| = q^n - \big|\mathcal{V}_{t,t,2s}(\mathbf{x})\big|.$$

In this setting, the main results[3] of Caro [36] and Wei [37] imply that G contains a clique of size at least

$$\sum_{\mathbf{x} \in \mathcal{B}_q(n)} \frac{1}{|\mathcal{B}_q(n)| - deg(\mathbf{x})} = \sum_{\mathbf{x} \in \mathcal{B}_q(n)} \frac{1}{|\mathcal{V}_{t,t,2s}(\mathbf{x})|}. \tag{13}$$

Recall from Lemma 5 that a clique of size k in G corresponds to a t-indel s-substitution correcting code of size k in $\mathcal{B}_q(n)$. Therefore, we conclude that there exists a t-indel s-substitution correcting code with a size as given by (13). Naturally, the size of this code forms a lower bound on $M_q(n,t,s)$. ∎

Given that Turán's theorem is implied by the result of Khogan [38], it is not surprising that Lemma 6 improves on Lemma 5. Indeed, using the convexity of $x \to \frac{1}{x}$ on $(0, \infty)$ it follows directly that

$$q^{-n} \sum_{\mathbf{x} \in \mathcal{B}_q(n)} \frac{1}{|\mathcal{V}_{t,t,2s}(\mathbf{x})|} \geq \frac{1}{q^{-n} \sum_{\mathbf{x} \in \mathcal{B}_q(n)} |\mathcal{V}_{t,t,2s}(\mathbf{x})|}. \tag{14}$$

In order to evaluate the lower bounds in Lemmas 5 & 6 the sizes of $V^{avg}_{t,t,2s}$ and $\mathcal{V}_{t,t,2s}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{B}_q(n)$ need to be determined, respectively. To the best of the authors' knowledge, analytic formulae for these expressions are not known for general parameters n, q, t and s. For this reason, we employ upper bounds on $V^{avg}_{t,t,2s}$ and $|\mathcal{V}_{t,t,2s}(\mathbf{x})|$ to obtain explicit and non-asymptotic results for both lemmas.

[3]To be precise, we translated the result from Caro and Wei on the existence of an independent set to the existence of a clique of the same size in the complement graph.

Theorem 1: For integers $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$ and $0 \leq s \leq n$, the following gives a lower bound on $M_q(n,t,s)$,

$$M_q(n,t,s) \geq \frac{q^{n+t}}{\left(\sum_{i=0}^{t} \binom{n}{i}(q-1)^i\right)^2 \sum_{i=0}^{2s} \binom{n-t}{i}(q-1)^i}. \tag{15}$$

Proof: We claim that $V^{avg}_{t,t,2s}$ can be upper bounded by

$$\frac{1}{q^t}\left(\sum_{i=0}^{t} \binom{n}{i}(q-1)^i\right)^2 \sum_{i=0}^{2s} \binom{n-t}{i}(q-1)^i. \tag{16}$$

In this case, the result of the theorem follows immediately from applying the upper bound to Lemma 5. Therefore, this proof is limited to proving the claim. In what follows, a superscript $^-$ will be used to denote a word in $\mathcal{B}_q(n-t)$, whereas an omission thereof is meant for words in $\mathcal{B}_q(n)$.

To this end, observe that each element in $\mathcal{V}_{t,t,2s}(\mathbf{x})$ can be reached from $\mathbf{x} \in \mathcal{B}_q(n)$ by first deleting precisely t symbols, followed by substituting at most $2s$ symbols and lastly inserting exactly t symbols. Hence, it follows that

$$|\mathcal{V}_{t,t,2s}(\mathbf{x})| \leq \sum_{\mathbf{y}^- \in \mathcal{D}_t(\mathbf{x})} \sum_{\mathbf{z}^- \in \mathcal{S}_{2s}(\mathbf{y}^-)} |\mathcal{I}_t(\mathbf{z}^-)|. \tag{17}$$

In order to evaluate the right-hand side of this expression, recall from (2) and (3) that the cardinalities of the sets $\mathcal{I}_t(\mathbf{x}^-)$ and $\mathcal{S}_{2s}(\mathbf{x}^-)$ do not depend on the choice of $\mathbf{x}^- \in \mathcal{B}_q(n-t)$. Moreover, the cardinality of $\mathcal{D}_t(\mathbf{x})$ averaged over all $\mathbf{x} \in \mathcal{B}_q(n)$ was given in (6). By combining these results and carefully taking into account the lengths of the words, it follows that

$$\begin{aligned}
V^{avg}_{t,t,2s} &= q^{-n} \sum_{\mathbf{x} \in \mathcal{B}_q(n)} |\mathcal{V}_{t,t,2s}(\mathbf{x})| \\
&\overset{(17)}{\leq} q^{-n} \sum_{\mathbf{x} \in \mathcal{B}_q(n)} \sum_{\mathbf{y}^- \in \mathcal{D}_t(\mathbf{x})} \sum_{\mathbf{z}^- \in \mathcal{S}_{2s}(\mathbf{y}^-)} |\mathcal{I}_t(\mathbf{z}^-)| \\
&\overset{(3)}{=} q^{-n} \sum_{\mathbf{x} \in \mathcal{B}_q(n)} \sum_{\mathbf{y}^- \in \mathcal{D}_t(\mathbf{x})} \sum_{\mathbf{z}^- \in \mathcal{S}_{2s}(\mathbf{y}^-)} S^t_{n,q} \\
&\overset{(2)}{=} q^{-n} \sum_{\mathbf{x} \in \mathcal{B}_q(n)} \sum_{\mathbf{y}^- \in \mathcal{D}_t(\mathbf{x})} S^{2s}_{n-t,q} \cdot S^t_{n,q} \\
&= q^{-n} \cdot S^t_{n,q} \cdot S^{2s}_{n-t,q} \cdot \sum_{\mathbf{x} \in \mathcal{B}_q(n)} |\mathcal{D}_t(\mathbf{x})| \\
&\overset{(6)}{=} q^{-t} \cdot \left(S^t_{n,q}\right)^2 \cdot S^{2s}_{n-t,q}.
\end{aligned}$$

Note that the last expression is equivalent to (16), which proves the claim. ∎

Observe that (9) and (10) are special cases of the latter theorem, since they are recovered by setting $t = 0$ and $s = 0$, respectively. Next, we derive an explicit lower bound on $M_q(n,t,s)$ using Lemma 6 and an upper bound on $|\mathcal{V}_{t,t,2s}(\mathbf{x})|$.

TABLE I
COMPARISON OF LOWER BOUNDS

n	$M_4(n,1,1)$ Thrm. 7	$M_4(n,1,1)$ Thrm. 8	$M_4(n,2,1)$ Thrm. 7	$M_4(n,2,1)$ Thrm. 8
8	2	3	1	1
10	13	13	1	1
12	93	96	2	2
14	783	802	10	9
16	7221	7371	63	58
18	71214	72533	475	442
20	740109	752444	3917	3671

Theorem 2: For integers $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$ and $0 \leq s \leq n$, the following gives a lower bound on $M_q(n,t,s)$,

$$M_q(n,t,s) \geq \sum_{r=1}^{n} \frac{q\binom{n-1}{r-1}(q-1)^{r-1}}{\binom{r+t-1}{t} \cdot \sum_{i=0}^{t}\binom{n}{i}(q-1)^i \cdot \sum_{j=0}^{2s}\binom{n-t}{j}(q-1)^j}.$$

Proof: Using similar reasoning as in the proof Theorem 1 it follows that

$$\begin{aligned}|\mathcal{V}_{t,t,2s}(\mathbf{x})| &\leq \sum_{\mathbf{y}^- \in \mathcal{D}_t(\mathbf{x})} \sum_{\mathbf{z}^- \in \mathcal{S}_{2s}(\mathbf{y}^-)} |\mathcal{I}_t(\mathbf{z}^-)| \\ &\leq \binom{r(\mathbf{x})+t-1}{t} \cdot S_{n,q}^{t} \cdot S_{n-t,q}^{2s},\end{aligned}$$

where we used (4), (2) and (3) for the cardinalities of $\mathcal{D}_t(\mathbf{x})$, $\mathcal{S}_{2s}(\mathbf{y}^-)$ and $\mathcal{I}_t(\mathbf{z}^-)$, respectively. Next, we apply these results to Lemma 6 and obtain the desired result,

$$\begin{aligned}M_q(n,t,s) &\geq \sum_{\mathbf{x}\in\mathcal{B}_q(n)} \frac{1}{|\mathcal{V}_{t,t,2s}(\mathbf{x})|} \\ &\geq \sum_{\mathbf{x}\in\mathcal{B}_q(n)} \frac{1}{\binom{r(\mathbf{x})+t-1}{t} \cdot S_{n,q}^{t} \cdot S_{n-t,q}^{2s}} \\ &= \sum_{r=1}^{n} \frac{q\binom{n-1}{r-1}(q-1)^{r-1}}{\binom{r+t-1}{t} \cdot S_{n,q}^{t} \cdot S_{n-t,q}^{2s}},\end{aligned}$$

where the last equality follows from (1). This chain of (in)equalities concludes the proof. ∎

Considering the implication of (14), it might be expected that Theorem 1 is also implied by Theorem 2. However, this is not the case in general because different bounds for $V_{t,t,2s}^{avg}$ and $|\mathcal{V}_{t,t,2s}(\mathbf{x})|$ were used in both theorems, respectively. Table I provides both instances in which Theorem 1 outperforms Theorem 2 and vice versa. Obviously, the bounds from Theorems 1 & 2 can be improved with the availability of exact expressions, or tighter bounds on $V_{t,t,2s}^{avg}$ and $|\mathcal{V}_{t,t,2s}(\mathbf{x})|$.

IV. THREE TYPES OF NON-ASYMPTOTIC UPPER BOUNDS

A. Singleton-Like Upper Bound

The well-known Singleton upper bound for s-substitution correcting codes [39] is given by $M_q(n,0,s) \leq q^{n-2s}$. Recently, Liu and Xing [18] proved a similar bound for t-indel correcting codes, $M_q(n,t,0) \leq q^{n-t}$. The following result combines these two bounds into an upper bound for t-indel s-substitution correcting codes.

Theorem 3: Let $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$ and $0 \leq s \leq n$ be integers such that $n - t - 2s \geq 0$. Then, the following holds,

$$M_q(n,t,s) \leq q^{n-t-2s}. \tag{18}$$

Proof: Let $\mathcal{C} \subseteq \mathcal{B}_q(n)$ be a t-indel s-substitution correcting code of maximal size. Consider the shortened code $\mathcal{C}^- \subseteq \mathcal{B}_q(n-t-2s)$ that is obtained from $\mathcal{C}$ by deleting the first $t + 2s$ symbols from all codewords in $\mathcal{C}$.

We claim that two distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$ yield two distinct codewords $\mathbf{c}_1^-, \mathbf{c}_2^- \in \mathcal{C}^-$. For contradiction, suppose that there exist two codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$ that agree on the last $n - t - 2s$ symbols. First, delete the first t symbols from both $\mathbf{c}_1$ and $\mathbf{c}_2$, to obtain $\mathbf{z}_1$ and $\mathbf{z}_2$, respectively. Notice that $\mathbf{z}_1$ and $\mathbf{z}_2$ agree on the last $n - 2s$ symbols. Hence, they differ in at most $2s$ places. This means that there exists some $\mathbf{z} \in \mathcal{B}_q(n-t)$ that can be obtained from both $\mathbf{z}_1$ and $\mathbf{z}_2$ by at most s substitutions. It follows that $\mathbf{z} \in \mathcal{S}_s(\mathbf{c}_1) \cap \mathcal{S}_s(\mathbf{c}_2)$ and in turn that $\mathbf{z} \in \mathcal{V}_{t,0,s}(\mathbf{c}_1) \cap \mathcal{V}_{t,0,s}(\mathbf{c}_2)$. This forms a contradiction with $\mathbf{c}_1$ and $\mathbf{c}_2$ being codewords of $\mathcal{C}$, because this intersection is empty according to Lemma 1. Therefore, the claim holds.

The claim implies that $\mathcal{C}$ and $\mathcal{C}^-$ have the same number of elements. It follows that $M_q(n,t,s) = |\mathcal{C}| = |\mathcal{C}^-| \leq q^{n-t-2s}$, since $\mathcal{C}$ was chosen to be maximal in the set of t-indel s-substitution correcting codes. The last chain of (in)equalities concludes the proof. ∎

B. Sphere-Packing Bounds

For s-substitution correcting codes the sphere-packing Hamming bound [11],

$$M_q(n,0,s) \leq \frac{q^n}{\sum_{i=0}^{s}\binom{n}{i}(q-1)^i},$$

is based on the idea that any such code $\mathcal{C}$ induces a disjoint set of spheres $\mathcal{S}_s(\mathbf{c})$ with $\mathbf{c} \in \mathcal{C}$. Naturally, the combined size of these disjoint spheres cannot exceed the size of $\mathcal{B}_q(n)$ which leads to the aforementioned bound.

For t-indel correcting codes, the same reasoning was used in [17] to show that

$$M_q(n,t,0) \leq \frac{q^{n+t}}{\sum_{i=0}^{t}\binom{n+t}{i}(q-1)^i},$$

by viewing these codes from the perspective of correcting solely insertions. From the perspective of correcting deletions only, an inattentive application of the sphere-packing argument leads to the Singleton bound. Namely, it holds that

$$M_q(n,t,0) \leq \frac{|\mathcal{B}_q(n-t)|}{\min_{\mathbf{x}\in\mathcal{B}_q(n)}|\mathcal{D}_t(\mathbf{x})|} = q^{n-t}, \tag{19}$$

where we used that the all-zero word $\mathbf{0} \in \mathcal{B}_q(n)$ satisfies $|\mathcal{D}_t(\mathbf{0})| = 1$. This bound can be improved by excluding the words for which $|\mathcal{D}_t(\mathbf{x})|$ is 'small' in the denominator of (19). Levenshtein [15] realized that these words are characterized by the words with few runs. For this reason, he partitioned the words in $\mathcal{B}_q(n)$ into two clusters based on their number of runs. Consequently, by applying the sphere-packing argument

only to words with many runs, i.e., for which $|\mathcal{D}_t(\mathbf{x})|$ is 'large', Levenshtein established for each $\max\{1, t-1\} \leq r \leq n$ that

$$M_q(n,t,0) \leq \frac{q^{n-t}}{\sum_{i=0}^{t}\binom{r+1-t}{i}} + q\sum_{i=1}^{r}\binom{n-1}{i-1}(q-1)^{i-1}.$$

In the context of t-indel s-substitution correcting codes, we will show that the sphere-packing argument can be used to obtain upper bounds for $M_q(n,t,s)$. Given that the sets $\mathcal{V}_{t',t'',s}(\mathbf{x})$ are also not of equal size for all $\mathbf{x} \in \mathcal{B}_q(n)$, the partitioning argument from Levenshtein based on the number of runs can be used in this setting as well. Obviously, this argument is not restricted to the use of only two clusters. The following lemma provides our sphere-packing bound in its most general form. Subsequently, this result will be made more concrete.

Lemma 7: Let $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$, $0 \leq s \leq n$ and $1 \leq k \leq n$ be integers. For each sequence of integers $0 = r_0 < r_1 < \cdots < r_k = n$, and for each pair of integers $0 \leq t', t'' \leq n$ such that $t' + t'' = t$, the following gives an upper bound on $M_q(n,t,s)$,

$$M_q(n,t,s) \leq \sum_{j=1}^{k} \frac{\sum_{r=a_j}^{b_j} q\binom{n-t'+t''-1}{r-1}(q-1)^{r-1}}{\min\limits_{\mathbf{x}\in\mathcal{A}_j} |\mathcal{V}_{t',t'',s}(\mathbf{x})|},$$

where for $1 \leq j \leq k$, $a_j \coloneqq \max\{1, r_{j-1}+1-2(t'+s)\}$, $b_j \coloneqq \min\{n-t'+t'', r_j + 2(t''+s)\}$ and $\mathcal{A}_j \coloneqq \{\mathbf{x} \in \mathcal{B}_q(n) : r_{j-1}+1 \leq r(\mathbf{x}) \leq r_j\}$.

Proof: Let $\mathcal{C} \subseteq \mathcal{B}_q(n)$ be a t-indel s-substitution correcting code of maximum size. The idea of this proof is to partition $\mathcal{B}_q(n)$ into k clusters based on the number of runs of the words in $\mathcal{B}_q(n)$. These clusters are given by $\mathcal{A}_j$ for $1 \leq j \leq k$. Since $\mathcal{C}$ is maximal and these clusters form a partition of $\mathcal{B}_q(n)$ it follows that $M_q(n,t,s) = \sum_{j=1}^{k} |\mathcal{C} \cap \mathcal{A}_j|$. Then, we bound each $|\mathcal{C} \cap \mathcal{A}_j|$ from above to arrive at the desired result.

Note that the clusters $\mathcal{A}_j$ indeed form a partition of $\mathcal{B}_q(n)$, because the sequence $0 = r_0 < r_1 < \cdots < r_k = n$ is strictly increasing. Let $1 \leq j \leq k$ and consider only the cluster $\mathcal{A}_j$. Let $\mathbf{c} \in \mathcal{C} \cap \mathcal{A}_j$ and consider a word $\mathbf{y} \in \mathcal{V}_{t',t'',s}(\mathbf{c})$. We claim that $a_j \leq r(\mathbf{y}) \leq b_j$ with a_j and b_j as given in the statement of this lemma. It holds that $\mathbf{y}$ has length $n - t' + t''$ and thus $1 \leq r(\mathbf{y}) \leq n - t' + t''$. Moreover, by Lemma 4 it follows that $r(\mathbf{c}) - 2(t'+s) \leq r(\mathbf{y}) \leq r(\mathbf{c}) + 2(t''+s)$, since $\mathbf{y} \in \mathcal{V}_{t',t'',s}(\mathbf{c})$. Together with $r_{j-1}+1 \leq r(\mathbf{c}) \leq r_j$ which follows from the definition of $\mathcal{A}_j$, we find that $r_{j-1}+1-2(t'+s) \leq r(\mathbf{y}) \leq r_j + 2(t''+s)$. Hence, we have proven the claim and continue with a sphere-packing argument.

The sets $\mathcal{V}_{t',t'',s}(\mathbf{c})$ with $\mathbf{c} \in \mathcal{C} \cap \mathcal{A}_j$ are disjoint according to Lemma 1, because $\mathcal{C}$ is a t-indel s-substitution correcting code. For this reason, the combined size of these spheres satisfies

$$\begin{aligned}\sum_{\mathbf{c}\in\mathcal{C}\cap\mathcal{A}_j} |\mathcal{V}_{t',t'',s}(\mathbf{c})| &= \left|\bigcup_{\mathbf{c}\in\mathcal{C}\cap\mathcal{A}_j} \mathcal{V}_{t',t'',s}(\mathbf{c})\right| \\ &\leq \left|\{\mathbf{y} \in \mathcal{B}_q(n-t'+t'') : a_j \leq r(\mathbf{y}) \leq b_j\}\right| \\ &\overset{(1)}{=} \sum_{r=a_j}^{b_j} q\binom{n-t'+t''-1}{r-1}(q-1)^{r-1},\end{aligned}$$

where we used the aforementioned claim to upper bound the size of the union. On the other hand, it also holds that

$$\begin{aligned}\sum_{\mathbf{c}\in\mathcal{C}\cap\mathcal{A}_j} |\mathcal{V}_{t',t'',s}(\mathbf{c})| &\geq |\mathcal{C}\cap\mathcal{A}_j| \cdot \min_{\mathbf{c}\in\mathcal{C}\cap\mathcal{A}_j} |\mathcal{V}_{t',t'',s}(\mathbf{c})| \\ &\geq |\mathcal{C}\cap\mathcal{A}_j| \cdot \min_{\mathbf{x}\in\mathcal{A}_j} |\mathcal{V}_{t',t'',s}(\mathbf{x})|,\end{aligned}$$

where we used that $\mathcal{C} \cap \mathcal{A}_j \subseteq \mathcal{A}_j$ in the last inequality. By combining the last two steps, it follows that

$$|\mathcal{C}\cap\mathcal{A}_j| \leq \frac{\sum_{r=a_j}^{b_j} q\binom{n-t'+t''-1}{r-1}(q-1)^{r-1}}{\min\limits_{\mathbf{x}\in\mathcal{A}_j} |\mathcal{V}_{t',t'',s}(\mathbf{x})|},$$

which concludes the proof. ∎

Intuitively, the parameters k and r_j for $1 \leq j \leq k$ in the previous lemma determine how $\mathcal{B}_q(n)$ is partitioned into k clusters based on the number of runs. Given the freedom of choice in these parameters, this lemma provides a family of upper bounds instead of a single bound. Note that $[a_j, b_j]$ for $1 \leq j \leq k$ are not pairwise disjoint, which affects the tightness of the bound. This results from the use of multiple clusters, which were introduced to benefit the tightness. In their current form, these bounds are implicit and can be made explicit by, e.g., (8) for $t = s = 1$.

Theorem 4: Let $n \geq 1$, $q \geq 2$ and $1 \leq k \leq n$ be integers. For each sequence of integers $0 = r_0 < r_1 < \cdots < r_k = n$, the following gives an upper bound on $M_q(n,1,1)$,

$$\begin{aligned}M_q(n,1,1) \leq{}& \frac{\sum_{r=a_1}^{b_1} q\binom{n-2}{r-1}(q-1)^{r-1}}{(n-1)(q-1)+1} \\ &+ \sum_{j=2}^{k} \frac{\sum_{r=a_j}^{b_j} q\binom{n-2}{r-1}(q-1)^{r-1}}{(r_{j-1}+1)((n-2)(q-1)-1)+q+2},\end{aligned}$$

with $a_j \coloneqq \max\{1, r_{j-1}-3\}$, $b_j \coloneqq \min\{n-1, r_j+2\}$ for $1 \leq j \leq k$.

Proof: Let $t' = 1$, $t'' = 0$ and $s = 1$. In this case, (8) yields

$$\min_{\mathbf{x}\in\mathcal{A}_1} |\mathcal{V}_{1,0,1}(\mathbf{x})| = (n-1)(q-1)+1,$$

and

$$\min_{\mathbf{x}\in\mathcal{A}_j} |\mathcal{V}_{1,0,1}(\mathbf{x})| = (r_{j-1}+1)((n-2)(q-1)-1)+q+2,$$

for $2 \leq j \leq k$ which follows because (8) is non-decreasing as function in r. Consequently, a direct application of Lemma 7 gives the desired result. ∎

For a general number of indels and substitutions, we apply the lower bound in Lemma 2 in order to make the family of bounds from Lemma 7 concrete.

Theorem 5: Let $n \geq 1$, $q \geq 2$, $0 \leq t \leq \frac{n}{2}$, $0 \leq s \leq \frac{n}{2}$ and $1 \leq k \leq n$ be integers. For each sequence of integers $0 = r_0 < r_1 < \cdots < r_k = n$, the following gives an upper bound on $M_q(n,t,s)$,

$$M_q(n,t,s) \leq \sum_{j=1}^{k} \frac{\sum_{r=a_j}^{b_j} q\binom{n-t-1}{r-1}(q-1)^{r-1}}{\sum_{i=0}^{s}\binom{\lfloor\frac{n}{2}\rfloor}{i}(q-1)^i \cdot \sum_{l=0}^{t}\binom{\lfloor\frac{r_{j-1}+1}{2}\rfloor}{l}},$$

with $a_j \coloneqq \max\{1, r_{j-1}+1-2(t+s)\}$, and $b_j \coloneqq \min\{n-t, r_j+2s\}$ for $1 \leq j \leq k$.

Proof: Let $t' = t$ and $t'' = 0$. In this case, Lemma 2 gives

$$\min_{\mathbf{x}\in\mathcal{A}_j} |\mathcal{V}_{t,0,s}(\mathbf{x})| \geq \sum_{i=0}^{s} \binom{\lfloor \frac{n}{2} \rfloor}{i} (q-1)^i \cdot \sum_{l=0}^{t} \binom{\lfloor \frac{r_{j-1}+1}{2} \rfloor - t}{l}$$

for $1 \leq j \leq k$. Here we used that the lower bound from Lemma 2 is non-decreasing as function in r. As a result, a direct application of Lemma 7 concludes this proof. ∎

C. Integer Linear Programming Bounds

At last, we employ an integer linear programming strategy to obtain non-asymptotic upper bounds on the maximum size of a code. This strategy that was initially described in [16] and used to derive various bounds on $M_q(n,t,0)$. The same strategy was used in [19] to derive upper bounds on $M_q(n,1,1)$ and $M_2(n,1,s)$. Here, we reconsider this strategy and derive several general upper bounds on $M_q(n,t,s)$ for $q \geq 2$ and t, $s \geq 0$, and argue that the bound for $M_q(n,1,1)$ from [19] can be improved.

In short, this strategy involves describing $M_q(n,t,s)$ as the optimal value of an integer linear program. An upper bound is then obtained by finding a feasible solution to the dual of a linear programming relaxation of this integer linear program. More specifically, this dual solution is given by any real-valued vector $\mathbf{w} = (w(\mathbf{y}))_{\mathbf{y}\in\mathcal{B}_q(n-t'+t'')}$ which satisfies the conditions: 1) $\mathbf{w} \geq \mathbf{0}$ and 2) $\sum_{\mathbf{y}\in\mathcal{V}_{t',t'',s}(\mathbf{x})} w(\mathbf{y}) \geq 1$ for all $\mathbf{x} \in \mathcal{B}_q(n)$. In turn, the upper bound is given by

$$M_q(n,t,s) \leq \sum_{\mathbf{y}\in\mathcal{B}_q(n-t'+t'')} w(\mathbf{y}). \tag{20}$$

A detailed description of this strategy can be found in [16] and is thus omitted here. All that remains to derive an explicit upper bound on $M_q(n,t,s)$ is to construct an appropriate vector $\mathbf{w}$. In a general and implicit form, this is accomplished by the next lemma.

Lemma 8: Let $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$ and $0 \leq s \leq n$ be integers. For each pair of integers $t', t'' \geq 0$ such that $t'+t'' = t$, and for each non-decreasing function $L : \{1,\ldots,n\} \to \mathbb{R}_{\geq 1}$ that satisfies $L(r(\mathbf{x})) \leq |\mathcal{V}_{t',t'',s}(\mathbf{x})|$ for all $\mathbf{x} \in \mathcal{B}_q(n)$, the following gives an upper bound on $M_q(n,t,s)$,

$$M_q(n,t,s) \leq \sum_{r=1}^{n-t'+t''} \frac{q\binom{n-t'+t''-1}{r-1}(q-1)^{r-1}}{L(c(r))},$$

where $c(r) = \max\{1, r-2(t''+s)\}$.

Proof: The aim of this proof is to construct a vector $\mathbf{w}$ that satisfies the aforementioned two conditions and then apply (20). To this end, define the vector $\mathbf{w}^*$ as follows

$$w^*(\mathbf{y}) = \frac{1}{L(c(r(\mathbf{y})))} = \begin{cases} \frac{1}{L(1)} & \text{if } r(\mathbf{y}) \leq 2(t''+s), \\ \frac{1}{L(r(\mathbf{y})-2(t''+s))} & \text{if } r(\mathbf{y}) \geq 2(t''+s), \end{cases}$$

for all $\mathbf{y} \in \mathcal{B}_q(n-t'+t'')$. We first show that $\mathbf{w}^*$ is well-defined and that $\mathbf{w}^*$ satisfies the two conditions.

The vector $\mathbf{w}^*$ is well-defined if $c(r(\mathbf{y})) \in \{1,\ldots,n\}$ for each $\mathbf{y} \in \mathcal{B}(n-t'+t'')$, because in that case $c(r(\mathbf{y}))$ is an element of the domain of L. Let $\mathbf{y} \in \mathcal{B}(n-t'+t'')$, then $c(r(\mathbf{y}))$ is integer-valued and clearly satisfies $c(r(\mathbf{y})) \geq 1$. Furthermore, it holds that

$$\begin{aligned} c(r(\mathbf{y})) &= \max\{1, r(\mathbf{y}) - 2\big(t''+s\big)\} \\ &\leq \max\{1, n-t'+t''-2\big(t''+s\big)\} \leq n. \end{aligned}$$

Hence, $\mathbf{w}^*$ is well-defined. The first condition $\mathbf{w}^* \geq \mathbf{0}$ is satisfied, because L is a strictly positive function. For the second condition, let $\mathbf{x} \in \mathcal{B}_q(n)$ and $\mathbf{y} \in \mathcal{V}_{t',t'',s}(\mathbf{x})$. Lemma 4 states that $r(\mathbf{y}) - 2(t''+s) \leq r(\mathbf{x})$, which gives

$$c(r(\mathbf{y})) = \max\{1, r(\mathbf{y}) - 2\big(t''+s\big)\} \leq r(\mathbf{x}).$$

This implies that $L(c(r(\mathbf{y}))) \leq L(r(\mathbf{x}))$, since L is a non-decreasing function, by definition. All in all, we obtain the second condition for $\mathbf{w}^*$,

$$\begin{aligned} \sum_{\mathbf{y}\in\mathcal{V}_{t',t'',s}(\mathbf{x})} w^*(\mathbf{y}) &= \sum_{\mathbf{y}\in\mathcal{V}_{t',t'',s}(\mathbf{x})} \frac{1}{L(c(r(\mathbf{y})))} \\ &\geq \sum_{\mathbf{y}\in\mathcal{V}_{t',t'',s}(\mathbf{x})} \frac{1}{L(r(\mathbf{x}))} \\ &\overset{(*)}{=} \frac{|\mathcal{V}_{t',t'',s}(\mathbf{x})|}{L(r(\mathbf{x}))} \overset{(**)}{\geq} 1, \end{aligned}$$

where we used in $(*)$ that the summands do not depend on $\mathbf{y}$, and in $(**)$ that $L(r(\mathbf{x}))$ bounds $|\mathcal{V}_{t',t'',s}(\mathbf{x})|$ from below. To conclude, $\mathbf{w}^*$ satisfies the two aforementioned conditions and an application of (20) yields the desired result,

$$\begin{aligned} M_q(n,t,s) &\leq \sum_{\mathbf{y}\in\mathcal{B}_q(n-t'+t')} w^*(\mathbf{y}) \\ &= \sum_{r=1}^{n-t'+t''} \sum_{\substack{\mathbf{y}\in\mathcal{B}_q(n-t'+t') \\ r(\mathbf{y})=r}} \frac{1}{L(c(r))} \\ &\overset{(1)}{=} \sum_{r=1}^{n-t'+t''} \frac{q\binom{n-t'+t''-1}{r-1}(q-1)^{r-1}}{L(c(r))}. \end{aligned}$$

The last chain of (in)equalities concludes the proof. ∎

The general formulation of Lemma 8 enables us to derive several explicit upper bounds on $M_q(n,t,s)$ using existing expressions and lower bounds on the size of $\mathcal{V}_{t',t'',s}(\mathbf{x})$.

For $t = s = 1$, let $L(r) = |\mathcal{V}_{1,0,1}(\mathbf{x})|$ for any $r \in \{1,\ldots,n\}$ and any $\mathbf{x} \in \mathcal{B}_q(n)$ with r runs. According to (8) the function L is well-defined and satisfies the conditions of Lemma 8. Therefore, the following upper bound is obtained,

$$\begin{aligned} M_q(n,1,1) \leq &\sum_{r=1}^{3} \frac{q\binom{n-2}{r-1}(q-1)^{r-1}}{(n-1)(q-1)+1} \\ &+ \sum_{r=4}^{n-1} \frac{q\binom{n-2}{r-1}(q-1)^{r-1}}{(r-2)((n-2)(q-1)-1)+q+2}. \end{aligned} \tag{21}$$

This bound was also obtained as an intermediate step in the proof of [19, Th. 4], but was further simplified to arrive at a cleaner expression, yet weaker bound. For this reason, (21) provides a stronger bound than [19, Th. 4].

TABLE II
COMPARISON OF UPPER BOUNDS ON $M_4(n,1,1)$

n	Singleton, Thrm. 9	Sphere-packing, Thrm. 11	ILP, (21)	ILP, [19, Thrm. 4]
4	**4**	6	6	-
6	64	**40**	**40**	361
8	1024	360	**278**	1112
10	16384	3184	**2331**	7509
12	262144	31436	**22723**	66596
14	4194304	334207	**244260**	677887
16	67108864	3758739	**2804591**	7508704

TABLE III
COMPARISON OF UPPER BOUNDS ON $M_4(n,2,2)$

n	Singleton, Thrm. 9	Sphere-packing, Thrm. 12	ILP, Thrm. 14	ILP, Thrm. 15
8	**16**	61	61	101
10	**256**	618	618	750
12	**4096**	6808	6553	6292
14	65536	79512	60576	**57827**
16	1048576	887433	**501328**	569001
18	16777216	7911135	**4035064**	5904749
20	268435456	64240727	**33533431**	63944612

For general t, s and $q \geq 2$, we employ the lower bound on $|\mathcal{V}_{t,0,s}(\mathbf{x})|$ from Lemma 2 and derive the following explicit upper bound.

Theorem 6: For integers $n \geq 1$, $q \geq 2$, $0 \leq t \leq \frac{n}{2}$ and $0 \leq s \leq \frac{n}{2}$, the following gives an upper bound on $M_q(n,t,s)$,

$$M_q(n,t,s) \leq \sum_{r=1}^{n-t} \frac{q\binom{n-t-1}{r-1}(q-1)^{r-1}}{\sum_{i=0}^{s}\binom{\lfloor \frac{n}{2} \rfloor}{i}(q-1)^i \cdot \sum_{j=0}^{t}\binom{\lfloor \frac{r-2s}{2} \rfloor - t}{j}}.$$

Proof: In the context of Lemma 8, let $t' = t$, $t'' = 0$ and let the lower bound L be given by Lemma 2. Observe that this lower bound is non-decreasing in r, and therefore it satisfies the conditions Lemma 8. As a result, Lemma 8 yields,

$$M_q(n,t,s) \leq \sum_{r=1}^{n-t} \frac{q\binom{n-t-1}{r-1}(q-1)^{r-1}}{\sum_{i=0}^{s}\binom{\lfloor \frac{n}{2} \rfloor}{i}(q-1)^i \cdot \sum_{j=0}^{t}\binom{\lfloor \frac{r-2s}{2} \rfloor - t}{j}},$$

which concludes the proof. ∎

Analogously, we also employ the lower bound on $|\mathcal{V}_{0,t,s}(\mathbf{x})|$ from Lemma 3 to derive a different bound.

Theorem 7: For integers $n \geq 1$, $q \geq 2$, $0 \leq t \leq \frac{n}{2}$ and $0 \leq s \leq \frac{n}{2}$, the following gives an upper bound on $M_q(n,t,s)$,

$$M_q(n,t,s) \leq \frac{q^{n+t}}{\sum_{i=0}^{s}\binom{\lfloor \frac{n}{2} \rfloor}{i}(q-1)^i \cdot \sum_{j=0}^{t}\binom{\lceil \frac{n}{2} \rceil + t}{j}(q-1)^j}. \tag{22}$$

Proof: In the context of Lemma 8, let $t' = 0$, $t'' = t$ and let the lower bound L be given by Lemma 3. Observe that L does not depend on $r(\mathbf{x})$, and thus trivially satisfies the conditions of the lower bound on $|\mathcal{V}_{0,t,s}(\mathbf{x})|$ in the statement of Lemma 8. As a result, Lemma 8 yields,

$$\begin{aligned} M_q(n,t,s) &\leq \sum_{r=1}^{n+t} \frac{q\binom{n+t-1}{r-1}(q-1)^{r-1}}{\sum_{i=0}^{s}\binom{\lfloor \frac{n}{2} \rfloor}{i}(q-1)^i \cdot \sum_{j=0}^{t}\binom{\lceil \frac{n}{2} \rceil + t}{j}(q-1)^j} \\ &= \frac{q^{n+t}}{\sum_{i=0}^{s}\binom{\lfloor \frac{n}{2} \rfloor}{i}(q-1)^i \cdot \sum_{j=0}^{t}\binom{\lceil \frac{n}{2} \rceil + t}{j}(q-1)^j}, \end{aligned}$$

where we used (1) in the last line. ∎

D. Comparison of Various Upper Bounds

Next, we numerically compare several of our upper bounds on $M_4(n,1,1)$ to an existing result in literature using Table II. Table III offers a similar comparison of the upper bounds on $M_4(n,2,2)$ in this paper. Note that the respective bounds have been rounded down to the nearest integer. The best bounds per row are indicated in bold. The columns for Theorem 4 & 5 display the best bound that is obtained after optimizing over all sequences $0 = r_0 < r_1 < \cdots < r_k = n$ with $k \leq 5$.

A key observation is that the best results in both tables are not given by a single bound. This shows that it has been worth deriving multiple bounds using different strategies. Despite its simplicity Theorem 3 is still relevant for yielding good results for small n. Table II also shows that the results in this paper provide improvements over the bound in [19, Th. 4]. The authors remark that numerical improvements can be made to the results of Table III by computing the cardinality of the set $\mathcal{V}_{t',t'',s}(\mathbf{x})$ numerically, and using these exact results instead of bounds on the size of this set. However, the purpose of this paper has been to derive several non-asymptotic bounds that hold for general q, t and s, and to a lesser extent to compute strong numerical bounds.

V. ASYMPTOTIC REDUNDANCY

In this section, we consider the setting in which the parameters q, t and s are fixed, and n tends to infinity. In this setting, Levenshtein [14] showed two asymptotic bounds on $M_2(n,t,s)$ which imply that the asymptotic redundancy of a binary t-indel s-substitution correcting code of maximal size lies between $(t+s)\log_2(n)$ and $(2t+2s)\log_2(n) + o(\log_2(n))$. Here, we provide alternative proofs for these asymptotic bounds and extend the results from binary to q-ary codes.

In what follows, we first derive the upper bound on the asymptotic redundancy by using Theorem 1.

Lemma 9: Let $q \geq 2$ be an integer. For non-negative integers s and t such that $s + t \geq 1$, the following holds

$$\limsup_{n\to\infty} \frac{n - \log_q\big(M_q(n,t,s)\big)}{(2t+2s)\log_q(n)} \leq 1.$$

Proof: Theorem 1 states that

$$M_q(n,t,s) \geq \frac{q^{n+t}}{\big(S^t_{n,q}\big)^2 \cdot S^{2s}_{n-t,q}}.$$

This implies that the redundancy of an optimal t-indel s-substitution correcting code is bounded by

$$n - \log_q(M_q(n,t,s)) \leq -t + 2\log_q\big(S^t_{n,q}\big) + \log_q\big(S^{2s}_{n-t,q}\big).$$

Note that for a fixed integer $k \geq 1$ it holds that $\binom{n}{k} = \frac{1}{k!}n^k + o(n^k)$. In turn, it follows that $S^s_{n,q} = \frac{(q-1)^s}{s!}n^s + o(n^s)$, and

$\log_q(S^s_{n,q}) = s\log_q(n) + o(\log_q(n))$. By combining these observations we obtain

$$\limsup_{n\to\infty} \frac{n - \log_q\big(M_q(n,t,s)\big)}{(2t+2s)\log_q(n)} \leq \limsup_{n\to\infty} \frac{-t + 2\log_q\big(S^t_{n,q}\big) + \log_q\Big(S^{2s}_{n-t,q}\Big)}{(2t+2s)\log_q(n)} = 1,$$

as desired. ∎

Next, we use Theorem 7 to derive the lower bound on the asymptotic redundancy.

Lemma 10: Let $q \geq 2$ be an integer. For non-negative integers s and t such that $s + t \geq 1$, the following holds

$$\liminf_{n\to\infty} \frac{n - \log_q\big(M_q(n,t,s)\big)}{(t+s)\log_q(n)} \geq 1.$$

Proof: For sufficiently large n, Theorem 7 states that

$$M_q(n,t,s) \leq \frac{q^{n+t}}{S^s_{\lfloor \frac{n}{2}\rfloor,q} \cdot S^t_{\lceil \frac{n}{2}\rceil+t,q}}.$$

This implies that the redundancy of an optimal t-indel s-substitution correcting code is bounded by

$$n - \log_q(M_q(n,t,s)) \geq -t + \log_q\Big(S^s_{\lfloor \frac{n}{2}\rfloor,q}\Big) + \log_q\Big(S^t_{\lceil \frac{n}{2}\rceil+t,q}\Big).$$

By similar reasoning as in the proof of Lemma 9 we obtain

$$\liminf_{n\to\infty} \frac{n - \log_q(M_q(n,t,s))}{(t+s)\log_q(n)} \geq \liminf_{n\to\infty} \frac{-t + \log_q\Big(S^s_{\lceil \frac{n}{2}\rceil,q}\Big) + \log_q\Big(S^t_{\lfloor \frac{n}{2}\rfloor+t,q}\Big)}{(t+s)\log_q(n)} = 1,$$

as desired. ∎

The following result is immediate from the last two lemmas.

Corollary 1: A maximal size t-indel s-substitution correcting code has an asymptotic redundancy that falls between $(t+s)\log_q(n) + o(\log_q(n))$ and $(2t+2s)\log_q(n) + o(\log_q(n))$.

VI. Concluding Remarks

In this paper, we have presented several non-asymptotic bounds on the maximal cardinality of t-indel s-substitution correcting codes. In order to improve these bounds, an interesting research challenge is to find an expression or tighter bounds for the cardinality of the set $\mathcal{V}_{t',t'',s}(\mathbf{x})$ for $\mathbf{x} \in \mathcal{B}_q(n)$.

In the context of DNA data storage, constrained coding techniques often include error-rate reduction measures, such as excluding long homopolymer runs, and requiring approximate GC/AT-balance. These measures naturally increase redundancy, but it is a priori unclear whether the reduction in error-rates outweighs the increase in redundancy. The results of this paper might act as a comparison whether including error-rate reduction measures is beneficial.

Appendix A
Proof of Lemma 2

Let $\mathbf{x}^1$ and $\mathbf{x}^2$ be the words which consist of the first $\lfloor \frac{n}{2} \rfloor$ symbols of $\mathbf{x}$, and the last $\lceil \frac{n}{2} \rceil$ symbols of $\mathbf{x}$, respectively. In other words, $\mathbf{x} = (\mathbf{x}^1|\mathbf{x}^2)$. Without loss of generality, we can assume that $\mathbf{x}^2$ contains at least $\lfloor \frac{r}{2} \rfloor$ runs. Otherwise, the order of the symbols in $\mathbf{x}$ can be reversed which does not affect the cardinality of $\mathcal{V}_{t,0,s}(\mathbf{x})$.

By concatenating two words $\mathbf{u} \in \mathcal{S}_s(\mathbf{x}^1)$ and $\mathbf{v} \in \mathcal{D}_t(\mathbf{x}^2)$, we obtain the word $(\mathbf{u}|\mathbf{v}) \in \mathcal{V}_{t,0,s}(\mathbf{x})$. Each such distinct pair of words $\mathbf{u}, \mathbf{v}$ yields a distinct word in $\mathcal{V}_{t,0,s}(\mathbf{x})$ and thus there exist at least $|\mathcal{S}_s(\mathbf{x}^1)| \cdot |\mathcal{D}_t(\mathbf{x}^2)|$ words in $\mathcal{V}_{t,0,s}(\mathbf{x})$. As a result,

$$|\mathcal{V}_{t,0,s}(\mathbf{x})| \geq \sum_{i=0}^{s} \binom{\lfloor \frac{n}{2} \rfloor}{i} (q-1)^i \cdot \sum_{j=0}^{t} \binom{\lfloor \frac{r}{2} \rfloor - t}{j},$$

where we used (2) for the cardinality of $\mathcal{S}_s(\mathbf{x}^1)$ and (5) as a lower bound on $|\mathcal{D}_t(\mathbf{x}^2)|$.

Appendix B
Proof of Lemma 3

Let $\mathbf{x}^1$ and $\mathbf{x}^2$ be the words which consist of the first $\lfloor \frac{n}{2} \rfloor$, and last $\lceil \frac{n}{2} \rceil$ symbols of $\mathbf{x}$, respectively. Using analogous reasoning as in the proof of Lemma 2, it follows that

$$|\mathcal{V}_{0,t,s}(\mathbf{x})| \geq |\mathcal{S}_s\left(\mathbf{x}^1\right)| \cdot |\mathcal{I}_t\left(\mathbf{x}^2\right)| = \sum_{i=0}^{s} \binom{\lceil \frac{n}{2} \rceil}{i} (q-1)^i \cdot \sum_{i=0}^{t} \binom{\lfloor \frac{n}{2} \rfloor + t}{i} (q-1)^i,$$

where we used (2) and (3) for the cardinality of $\mathcal{S}_s(\mathbf{x}^1)$ and $\mathcal{I}_t(\mathbf{x}^2)$, respectively.

Appendix C
Proof of Lemma 4

Consider an arbitrary word $\mathbf{x} \in \mathcal{B}_q(n)$ with $r = r(\mathbf{x})$ runs, then we argue how a single deletion, insertion or substitution in $\mathbf{x}$ can affect the number of runs of $\mathbf{x}$. Let l_j denote the length of the j-th run in $\mathbf{x}$.

After a single deletion in the i-th run of $\mathbf{x}$, we claim that the number of runs in $\mathbf{x}$ decreases by at most two, but does not increase. We consider four cases. First, in case $l_i > 1$, then the number of runs is unchanged because the i-th run is shortened by one, and no runs are deleted or created. In case $l_i = 1$ and the i-th run is the first or last run in $\mathbf{x}$, then removing this run does not affect the other runs and reduces the number of runs by one. In case, $l_i = 1$ and the $(i-1)$-th and $(i+1)$-th run contain the same symbol values, then the number of runs decreases by two. Lastly, in case $l_i = 1$, and the $(i-1)$-th and $(i+1)$-th run contain different symbols, then the number of runs decreases by one, because only the i-th run is removed. The four cases jointly prove the claim.

As a result of a single insertion into $\mathbf{x}$, the number of runs in $\mathbf{x}$ cannot decrease and can increase by at most two. This follows from the previous claim combined with the fact that any insertion can be reversed by a deletion, and vice versa.

For a single substitution we claim that the number of runs in $\mathbf{x}$ can both be increased and decreased by at most two. We consider three cases. First, suppose that a substitution is carried out in a unit run. Then it can either remain a unit run in which case the number of runs is unaffected, or it joins with one or two neighboring runs in which case the number of runs decreases by one or two, respectively. Secondly, assume the substitution is performed in the first or last element of a run of length at least two, then either a unit run is formed in

the position of the substitution or the symbol is changed such that it matches the symbols in the neighboring run. Hence, the number of runs increases by one or does not change. Lastly, suppose that the substitution is carried out in a run of length at least three and neither in the first nor last element of this run. In this case, a unit run is created at the position of the substitution as well as two runs on either side of it. Therefore, the number of runs increases by precisely two. Combining the results from these three cases yields the claim.

By repeatedly applying single edits, the previous reasoning shows that after t' deletions, t'' insertions and s substitutions the number of runs in $\mathbf{x}$ decreases by at most $2(t' + s)$ and increases by at most $2(t'' + s)$.

REFERENCES

[1] G. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, p. 1628, Sep. 2012.

[2] R. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angew. Chem. Int. Ed. Engl.*, vol. 54, pp. 2552–2555, Feb. 2015.

[3] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable random-access DNA-based storage system," *Sci. Rep.*, vol. 5, Sep. 2015, Art. no. 14138.

[4] S. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Sci. Rep.*, vol. 7, p. 5011, Jul. 2017.

[5] P. Antkowiak et al., "Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction," *Nat. Commun.*, vol. 11, p. 5345, Oct. 2020.

[6] S. M. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *IEEE Trans. Mol., Biol. Multi-Scale Commun.*, vol. 1, no. 3, pp. 230–248, Sep. 2015.

[7] M. G. Ross et al., "Characterizing and measuring bias in sequence data," *Genome Biol.*, vol. 14, no. 5, p. 51, May 2013.

[8] R. Roth, *Introduction to Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[9] E. N. Gilbert, "A comparison of signalling alphabets," *Bell Syst. Tech. J.*, vol. 31, no. 3, pp. 504–522, May 1952.

[10] R. R. Varshamov, "Estimate of the number of signals in error correcting codes," *Doklady Akademii Nauk SSSR*, vol. 117, no. 5, pp. 739–741, Jun. 1957.

[11] R. W. Hamming, "Error detecting and error correcting codes," *Bell Syst. Tech. J.*, vol. 29, no. 2, pp. 147–160, Apr. 1950.

[12] M. J. E. Golay, "Notes on digital coding," *Proc. IRE*, vol. 37, p. 657, Jun. 1949.

[13] P. Delsarte, "Bounds for unrestricted codes, by linear programming," *Philips Res. Rep.*, vol. 27, pp. 272–289, May 1972.

[14] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, Feb. 1966.

[15] V. I. Levenshtein, "Bounds for deletion/insertion correcting codes," *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2002, p. 370.

[16] A. A. Kulkarni and N. Kiyavash, "Non-asymptotic upper bounds for deletion correcting codes," *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 5115–5130, Apr. 2013.

[17] L. Tolhuizen, "Upper bounds on the size of insertion/deletion correcting codes," in *Proc. 8th Int. Workshop Algebraic Combinatorial Coding Theory*, Sep. 2002, pp. 242–246.

[18] S. Liu and C. Xing, "Bounds and constructions for insertion and deletion codes," *IEEE Trans. Inf. Theory*, vol. 69, no. 2, pp. 928–940, Feb. 2023.

[19] I. Smagloy, L. Welter, A. Wachter-Zeh, and E. Yaakobi, "Single-deletion single-substitution correcting codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aug. 2020, pp. 775–780.

[20] W. Song, N. Polyanskii, K. Cai, and X. He, "On multiple-deletion multiple-substitution correcting codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Sep. 2021, pp. 2655–2660.

[21] W. Song, N. Polyanskii, K. Cai, and X. He, "Systematic codes correcting multiple-deletion and multiple-substitution errors," *IEEE Trans. Inf. Theory*, vol. 68, no. 10, pp. 6402–6416, Oct. 2022.

[22] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Sci. Rep.*, vol. 9, p. 9663, Jul. 2019.

[23] D. Cullina and N. Kiyavash, "An improvement to Levenshtein's upper bound on the cardinality of deletion correcting codes," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3862–3870, Jul. 2014.

[24] M. Abu-Sini and E. Yaakobi, "On Levenshtein's reconstruction problem under insertions, deletions, and substitutions," *IEEE Trans. Inf. Theory*, vol. 67, no. 11, pp. 7132–7158, Nov. 2021.

[25] V. I. Levenshtein, "Elements of the coding theory (in Russian)," in *Proc. Discr. Math. Mathe. Problems Cybern.*, 1974, pp. 207–235.

[26] D. S. Hirschberg and M. Regnier, "Tight bounds on the number of string subsequences," *J. Discr. Algorithms*, vol. 1, no. 1, pp. 123–132, Jun. 2001.

[27] Y. Liron and M. Langberg, "A characterization of the number of subsequences obtained via the deletion channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2300–2312, May 2015.

[28] H. Mercier, M. Khabbazian, and V. K. Bhargava, "On the number of subsequences when deleting symbols from a string," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3279–3285, Jun. 2008.

[29] M. Abu-Sini and E. Yaakobi, "Reconstruction of sequences in DNA storage," *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Sep. 2019, pp. 290–294.

[30] F. Sala and L. Dolecek, "Counting sequences obtained from the synchronization channel," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 2925–2929.

[31] L. Tolhuizen, "The generalized Gilbert-Varshamov bound is implied by Turan's theorem," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1605–1606, Sep. 1997.

[32] J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[33] F. Sala, R. Gabrys, and L. Dolecek, "Gilbert-Varshamov-like lower bounds for deletion-correcting codes," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2014, pp. 147–151.

[34] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Three novel combinatorial theorems for the insertion/deletion channel," *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 2702–2706.

[35] J. Gu and T. Fuja, "A generalized Gilbert-Varshamov bound derived via analysis of a code-search algorithm," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 1089–1093, May 1993.

[36] Y. Caro, *New Results on the Independence Number*, Tel-Aviv Univ., Tel Aviv, Israel, pp. 75–79, 1979.

[37] V. Wei, "A lower bound on the stability number of a simple graph," in *Bell Laboratories Technical Memorandum*. New Providence, NJ, USA: Murray Hill, 1981.

[38] S. Khogan, "A note on a Caro–Wei bound for the bipartite independence number in graphs," *Discr. Math.*, vol. 344, no. 4, Apr. 2021, Art. no. 112285.

[39] R. Singleton, "Maximum distance q-nary codes," *IEEE Trans. Inf. Theory*, vol. 10, no. 2, pp. 116–118, Apr. 1964.