On the Duplication Distance of Binary Strings

Noga Alon
Schools of Mathematics and Computer Science
Tel Aviv University
Tel Aviv 6997801, Israel
Email: nogaa@post.tau.ac.il

Jehoshua Bruck, Farzad Farnoud, and Siddharth Jain
Department of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, USA
Email: {bruck,farnoud,sidjain}@caltech.edu

***Abstract*—We study the tandem duplication distance between binary sequences and their roots. This distance is motivated by genomic tandem duplication mutations and counts the smallest number of tandem duplication events that are required to take one sequence to another. We consider both exact and approximate tandem duplications, the latter leading to a combined duplication/Hamming distance. The paper focuses on the maximum value of the duplication distance to the root. For exact duplication, denoting the maximum distance to the root of a sequence of length n by $f(n)$, we prove that $f(n) = \Theta(n)$. For the case of approximate duplication, where a β-fraction of symbols may be duplicated incorrectly, we show using the Plotkin bound that the maximum distance has a sharp transition from linear to logarithmic in n at $\beta = 1/2$.**

I. Introduction

The genome of every organism is subject to mutations resulting from imperfect genome replication, as well as environmental factors. These mutations include *tandem duplications*, which create *tandem repeats* by duplicating a substring and inserting it adjacent to the original; *point mutations*, which substitute a base in the sequence by another; and *deletions*, in which a substring is removed from the sequence.

Gaining a better understanding of mutations that modify genomes – thereby creating the variety needed for natural selection – is helpful in many fields including phylogenomics, systems biology, medicine, and bioinformatics. One aspect of this task, namely studying the ability of duplication mutations to generate diversity, has been recently studied from an information-theoretic point of view [1], [2], [3]. In particular, [1] models sequences generated from a starting "seed" through different types of duplication as string systems and studies their *capacity* and *expressiveness*. The notion of capacity quantifies the ability of the systems to generate diverse families of sequences, and expressiveness is concerned with determining whether every sequence can be generated as a substring of another sequence, if not independently. The results in [1], [2] include lower bounds on the capacity of tandem duplications and establishing that certain systems have nonzero capacity.

The aforementioned works focus on the possibility of generating sequences and do not consider the number of duplication steps it takes to do so for any given sequence, which is the subject of the current paper. Specifically, we define distance measures between pairs of sequences based on the number of exact or approximate tandem duplications that are needed to transform one sequence to the other. We then study the distances between sequences of length $n \in \mathbb{N}$ and their roots, i.e., the shortest sequences from which they can be obtained via these operations.

From an evolutionary point of view, the duplication distance between a sequence and its root is of interest since it corresponds to a likely path through which a root may have evolved into the sequence under study, especially in DNA tandem repeat regions, which form about 3% of the human genome [4], assuming that mutations are unlikely events. The search for such a path has been studied in the literature, e.g., in [5], [6]. These works, however, have a more restrictive duplication model than that of the present paper. Furthermore, we are focused on the extremal distance values, while they study the problem from an algorithmic point of view.

Formally, a (*tandem*) *repeat of length* h in a sequence is two identical adjacent blocks, each consisting of h consecutive elements. For example, the sequence $12\underline{134134}51$ contains the repeat 134134 of length 3. A repeat of length h may be created through a duplication of length h and removed through a *deduplication* of length h, i.e., by removing one of the two adjacent identical blocks. The *duplication/deduplication distance* between two sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ is the smallest number of duplications and deduplications that can turn $\boldsymbol{x}$ into $\boldsymbol{y}$ (to denote sequences we use bold symbols.). We set the distance to infinity if the task is not possible, for example, if $\boldsymbol{x} = 1$ and $\boldsymbol{y} = 0$.

For two sequences $\boldsymbol{x}$ and $\boldsymbol{y}$, if $\boldsymbol{y}$ can be obtained from $\boldsymbol{x}$ through duplications, we say that $\boldsymbol{x}$ is an *ancestor* of $\boldsymbol{y}$ and that $\boldsymbol{y}$ is a *descendant* of $\boldsymbol{x}$. An ancestor $\boldsymbol{x}$ of $\boldsymbol{y}$ is a *root* of $\boldsymbol{y}$ if it is *square-free*, i.e., it does not contain a repeat. We define the *duplication distance* between two sequences as the minimum number of duplications required to convert the shorter sequence to the longer one.[1] This distance is finite if and only if one sequence is the ancestor of the other.

This paper is focused on finding bounds on the duplication distance of binary sequences to their roots. Our attention is limited to binary sequences for the sake of simplicity, since for the binary alphabet, the root of every sequence is unique and belongs to the set $\{0, 1, 01, 10, 010, 101\}$. Specifically, the roots of 0^n and 1^n are 0 and 1, respectively. For every other binary sequence $\boldsymbol{s}$ of length n, the root of $\boldsymbol{s}$ is the sequence

[1] Note that using the term distance here is a slight abuse of notation as the duplication distance does not satisfy the triangle inequality.

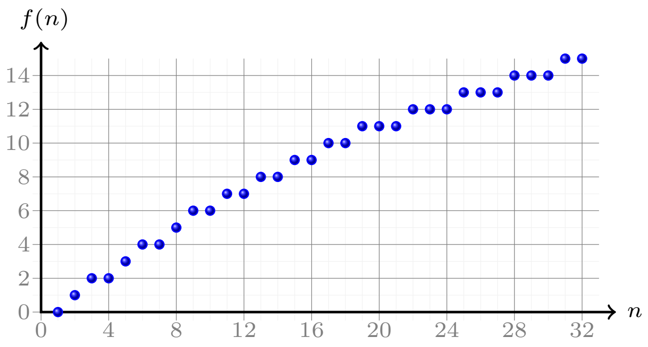

Fig. 1. $f(n)$ for $1 \leq n \leq 32$ obtained by computer search.

in the set $\{01, 10, 010, 101\}$ that starts and ends with the same symbols as $\boldsymbol{s}$. For example, the root of $\boldsymbol{s} = 1001011$ is 101 since

$$\boldsymbol{s} = 1\underline{00}1011 \xrightarrow{dd} 1010\underline{11} \xrightarrow{dd} \underline{1010}1 \xrightarrow{dd} 101,$$

where $\boldsymbol{x} \xrightarrow{dd} \boldsymbol{y}$ indicates that $\boldsymbol{y}$ can be obtained from $\boldsymbol{x}$ through a deduplication. More generally, through deduplication, we can convert every binary sequence to its root by first converting every run of 1s or 0s to one symbol and then converting a run of 10s to 10 or a run of 01s to 01.

We note that a celebrated result by Thue from 1906 [7] states that for alphabets of size ≥ 3, there is an infinite square-free sequence. Thus the set of roots for such alphabets is infinite (in contrast to the binary alphabet).

For a binary sequence $\boldsymbol{s}$, let $f(\boldsymbol{s})$ denote the duplication distance between $\boldsymbol{s}$ and its root and let $f(n)$ be the maximum of $f(\boldsymbol{s})$ for all sequences $\boldsymbol{s}$ of length n. Figure 1, which was obtained through computer search, shows the values of $f(n)$ for $1 \leq n \leq 32$. In this paper, we provide bounds on $f(\boldsymbol{s})$ and on $f(n)$. For example, we show that $f(n) = \Omega(n)$ by encoding every sequences using its root and a specific sequence of duplications that generate it when applied to its root in order, such that the number of valid encodings is $2^{O(f(n))}$. The desired result follows from the fact that we must have $2^{O(f(n))} \geq 2^n$. On the other hand, we show constructively that $\lim_n f(n)/n \leq 17/32$.

We also consider a variation of the duplication distance, referred to as the *duplication/Hamming distance*, where the duplication process is imprecise and so the inserted block is not necessarily an exact copy. More precisely, the *β-duplication/Hamming distance* between two sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ is the smallest number of duplications that can turn the shorter sequence into the longer one, where each duplication may produce a block that differs from the original in at most a β-fraction of positions. The shortest distance between $\boldsymbol{s}$ and any of its roots is denoted by $f_\beta(\boldsymbol{s})$ and the maximum of $f_\beta(\boldsymbol{s})$ over all sequences $\boldsymbol{s}$ of length n is denoted by $f_\beta(n)$.

While it is clear that $f_\beta(n) \leq f(n)$, by extending our encoding method used to obtain a lower bound on $f(n)$, we show that $f_\beta(n) = \Omega(n)$ for *any* constant $\beta < 1/2$. Furthermore, we show that there is a sharp transition at $\beta = 1/2$: For $\beta > 1/2$, we have $f_\beta(n) = O(\log n)$. The proof of this statement relies on viewing certain substrings of any sequence $\boldsymbol{s}$ as a code and then using the Plotkin bound [8] to show that the minimum distance of this code is sufficiently small to ensure the existence of adjacent blocks whose Hamming distance is at most a β-fraction of their length.

The rest of the paper is organized as follows. In the next section, we consider the duplication distance and present the bounds on $f(n)$. Section III considers the duplication/Hamming distance. We conclude the paper in Section IV with some open problems.

II. Duplication Distance

We start this section by providing straight-forward bounds on $f(\boldsymbol{s})$ for a sequence $\boldsymbol{s}$. Suppose the root of $\boldsymbol{s}$ is $\boldsymbol{\sigma} \in \{0, 1, 01, 10, 010, 101\}$. Clearly,

$$\log \frac{|\boldsymbol{s}|}{|\boldsymbol{\sigma}|} \leq f(\boldsymbol{s}) \leq |\boldsymbol{s}|,$$

where the log in the equation, as well as all others in the paper, is in base 2, and $|\boldsymbol{x}|$ denotes the length of $\boldsymbol{x}$. While the above lower bound is tight, in the sense that there exist arbitrarily long sequences $\boldsymbol{s}$ that satisfy it with equality, e.g., $\boldsymbol{s} = 0^{2^k}$ and $\boldsymbol{\sigma} = 0$, the upper bound is not tight as we will see.

We now provide a lower bound on $f(\boldsymbol{s})$ that depends on the number of distinct k-mers (substrings of length k) of $\boldsymbol{s}$, denoted $K(\boldsymbol{s})$, for a positive integer value of k.

Lemma 1. *For a sequence $\boldsymbol{s}$ and a positive integer $k \geq 4$,*

$$f(\boldsymbol{s}) \geq \frac{K(\boldsymbol{s})}{k-1} .$$

Proof. For two sequences $\boldsymbol{x} = \boldsymbol{tuuv}$ and $\boldsymbol{y} = \boldsymbol{tuv}$, we have $K(\boldsymbol{y}) \geq K(\boldsymbol{x}) - (k-1)$, since the only case in which a k-mer occurs in $\boldsymbol{x}$ but not in $\boldsymbol{y}$ is when the only instance of that k-mer intersects both copies of $\boldsymbol{u}$ in $\boldsymbol{x}$. There are at most $k-1$ k-substrings of $\boldsymbol{x}$ that intersect both copies of $\boldsymbol{u}$. Finally, no root contains a k-mer for $k \geq 4$. □

An immediate corollary of the lemma provides a construction for sequences whose duplication distance from the root is $f(\boldsymbol{s}) = \Omega(n/\log n)$. This is the largest distance for which we have a construction, although we will later show that sequences with $f(\boldsymbol{s}) = \Omega(n)$ exist.

Corollary 2. *For any binary De Bruijn sequence $\boldsymbol{s}$ of order k (which has length $n = 2^k$), we have*

$$f(\boldsymbol{s}) \geq \frac{n - \log n}{\log n} .$$

It is worth noting that using the same technique as the proof of $f(n) = \Omega(n)$ in Theorem 3, and the fact that there are $\frac{2^{n/2}}{n}$ De Bruijn sequences of length n when n is a power of two, one can show that the largest deduplication distance for these sequences grows linearly in n.

Next we present one of the main results of the paper which shows that $f(n) = \Theta(n)$ and that for almost all sequences $\boldsymbol{s}$, $f(\boldsymbol{s})$ increases linearly with n.

Theorem 3. *The limit* $\lim_{n\to\infty} f(n)/n$ *exists and*

$$0.045 \le \lim_{n\to\infty} \frac{f(n)}{n} \le \frac{17}{32} \cdot$$

Furthermore, for sufficiently large n*,* $f(\boldsymbol{s}) \ge 0.045n$ *for all but an exponentially small fraction of sequences* $\boldsymbol{s}$ *of length* n*.*

The lower bound in Theorem 3 is proved with the help of Theorem 4, and its limit statement and the upper bound are proved using Theorem 7.

Theorem 4. *For* $0 < \alpha < 1$*, consider the set of the* $\lfloor 2^{n\alpha} \rfloor$ *sequences of length* n *with the smallest duplication distance to the root and let* F_α *be the maximum duplication distance to the root for a sequence in this set. Then*

$$6n \sum_{f=1}^{F_\alpha} \binom{n+f}{f} \binom{2n+f}{f} \binom{2n+f+2}{f} 2^f \ge 2^{n\alpha} - 1. \quad (1)$$

Before stating the proof, we present some background, definitions, and a claim that will be useful in the proof, as well as a simpler but weaker result.

Recall that if the sequence $\boldsymbol{s} = s_1 s_2 \cdots s_m$ contains a repeat, then omitting one of the two blocks of this repeat to obtain a new sequence is called a deduplication. We also refer to the resulting sequence $\boldsymbol{s}'$ as a deduplication of $\boldsymbol{s}$, and write $\boldsymbol{s} \xrightarrow{dd} \boldsymbol{s}'$. A *deduplication process* for a binary sequence $\boldsymbol{s}$ is a sequence of sequences $\boldsymbol{s} = \boldsymbol{s}_0 \xrightarrow{dd} \boldsymbol{s}_1 \xrightarrow{dd} \boldsymbol{s}_2 \xrightarrow{dd} \cdots \xrightarrow{dd} \boldsymbol{s}_f$, where each $\boldsymbol{s}_{i+1}$ is a deduplication of $\boldsymbol{s}_i$ and the final sequence $\boldsymbol{s}_f$ is the (square-free) root of $\boldsymbol{s}$. The *length* of the deduplication process above is f, that is, the number of deduplications in it. A deduplication of $\boldsymbol{s}$ is an (i,h)-*step* if i is the starting position of (the first block) of a repeat of length h and one of the blocks of this repeat is omitted. For example, if $\boldsymbol{s} = 123\underline{134}13451$, a $(4,3)$-step produces $\boldsymbol{s}' = 12313451$. A deduplication process of length f of a sequence $\boldsymbol{s}$ can be described by a sequence of pairs $(i_t, h_t)_{t=1}^{f}$, where step number t is an (i_t, h_t)-step. It is not difficult to check that knowing the final sequence in the process, and knowing all the pairs (i_t, h_t) of deduplications in the process, in order, we can reconstruct the original sequence $\boldsymbol{s}$.

From the preceding discussion, each binary sequence $\boldsymbol{s}$ can be encoded as the pair $\left(\boldsymbol{\sigma}, (i_t, h_t)_{t=1}^{f(\boldsymbol{s})}\right)$, where $\boldsymbol{\sigma}$ is the root of $\boldsymbol{s}$ and $(i_t, h_t)_{t=1}^{f(\boldsymbol{s})}$ a deduplication process of $\boldsymbol{s}$. Since there are only 6 possibilities for $\boldsymbol{\sigma}$, and less than n^2 possibilities for each pair (i_t, h_t), if $F = f(n)$, then

$$6 \sum_{f=1}^{F} \left(n^2\right)^f \ge 2^n, \quad (2)$$

which implies that $F = f(n) = \Omega(n/\log n)$.

In the aforementioned encoding, several deduplication processes may map to the same sequence. We improve upon (2) by defining deduplication processes of a special form that remove some of the redundancy, and by doing so, we obtain (1), which will lead to the linear lower bound of Theorem 3.

Definition 5. A deduplication process $\boldsymbol{s} = \boldsymbol{s}_0 \xrightarrow{dd} \boldsymbol{s}_1 \xrightarrow{dd} \boldsymbol{s}_2 \xrightarrow{dd} \cdots \xrightarrow{dd} \boldsymbol{s}_f$ of a sequence $\boldsymbol{s}$, in which the steps are $(i_1, h_1), (i_2, h_2), \ldots, (i_f, h_f)$, is *normal* if the following condition holds: For any $1 \le t < f$, if $i_{t+1} < i_t$ then $i_{t+1} + 2h_{t+1} \ge i_t$.

The following claim shows that if we limit ourselves to normal deduplication processes, we can still encode every binary sequence with processes of the same length.

Claim 6. *For any deduplication process* $\boldsymbol{s} = \boldsymbol{s}_0 \xrightarrow{dd} \boldsymbol{s}_1 \xrightarrow{dd} \boldsymbol{s}_2 \xrightarrow{dd} \cdots \xrightarrow{dd} \boldsymbol{s}_f$ *of length* f *of a sequence* $\boldsymbol{s}$*, there is a normal deduplication process* $\boldsymbol{s} = \boldsymbol{s}_0 \xrightarrow{dd} \boldsymbol{s}'_1 \xrightarrow{dd} \boldsymbol{s}'_2 \xrightarrow{dd} \cdots \xrightarrow{dd} \boldsymbol{s}'_f = \boldsymbol{s}_f$ *of the same length, with the same final sequence.*

Proof. Among all deduplication processes of length f starting with $\boldsymbol{s}$ and ending with $\boldsymbol{s}_f$, consider the one minimizing the number of pairs (i_t, h_t), (i_q, h_q) with $1 \le t < q \le f$, and $i_q < i_t$. We claim that this process is normal. Indeed, otherwise there is some t, $1 \le t < f$ so that $i_{t+1} < i_t$ and $i_{t+1} + 2h_{t+1} < i_t$. But in this case we can switch the steps (i_t, h_t) and (i_{t+1}, h_{t+1}), performing the step (i_{t+1}, h_{t+1}) just before (i_t, h_t). This will clearly leave all sequences $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_f$, besides $\boldsymbol{s}_t$, the same, and in particular $\boldsymbol{s}_0 = \boldsymbol{s}$ and $\boldsymbol{s}_f$ stay the same. This contradicts the minimality in the choice of the process, establishing the claim. □

We now turn to the proof of Theorem 4.

Proof of Theorem 4. Let U_α denote the set of $\lfloor 2^{n\alpha} \rfloor$ sequences that have the smallest duplication distances to their roots among binary sequences of length n and recall that $F_\alpha = \max\{f(\boldsymbol{s}) : \boldsymbol{s} \in U_\alpha\}$. By Claim 6, for each of the sequences $\boldsymbol{s}$ of U_α, there is a normal deduplication process $\boldsymbol{s} = \boldsymbol{s}_0 \xrightarrow{dd} \boldsymbol{s}_1 \xrightarrow{dd} \boldsymbol{s}_2 \xrightarrow{dd} \cdots \xrightarrow{dd} \boldsymbol{s}_f$ of length $f \le F_\alpha$. Let the steps of this process be $(i_1, h_1), (i_2, h_2), \ldots, (i_f, h_f)$. As before, it is clear that knowing the final sequence $\boldsymbol{s}_f$ and all the pairs (i_t, h_t), we can reconstruct $\boldsymbol{s}$. There are 6 possibilities for $\boldsymbol{s}_f$. As each step (i_t, h_t) reduces the length of the sequence by h_t it follows that $\sum_{i=1}^{f} h_t < n$ and therefore there are at most $\binom{n+f}{f}$ possibilities for the sequence $(h_1, h_2, h_3, \ldots, h_f)$. In order to record the sequence $(i_1, i_2, \ldots, i_f)$ it suffices to record i_1 and all the differences $i_t - i_{t+1}$ for all $1 \le t < n$. There are less than n possibilities for i_1, and there are at most 2^f possibilities for deciding about the set of all indices t for which the difference $i_t - i_{t+1}$ is positive. As the process is normal, for each such positive difference, we know that $i_{t+1} + 2h_{t+1} \ge i_t$, that is $i_t - i_{t+1} \le 2h_{t+1}$. It follows that the sum of all positive differences, $\sum_{t: i_t - i_{t+1} > 0} (i_t - i_{t+1})$, is at most $2\sum_t h_t < 2n$, and hence the number of choices for these differences is at most $\binom{2n+f}{f}$.

Since $i_f \le 3$, we have $i_1 - i_f \ge 1 - 3 = -2$. So

$$\sum_{t: i_t - i_{t+1} \le 0} (i_t - i_{t+1}) \ge -2 - \sum_{t: i_t - i_{t+1} > 0} (i_t - i_{t+1}) > -2 - 2n.$$

Therefore, the number of choices for all non-positive differences $i_t - i_{t+1}$ is at most $\binom{2n+f+2}{f}$. Putting all of these together, and noting that $|U_\alpha| \geq 2^{n\alpha}-1$, implies the assertion of Theorem 4. □

Since $\binom{p}{q} \leq 2^{pH(q/p)}$ for positive integers $0 < q < p$ [8, p. 309], Theorem 4 implies that

$$3\left(2+\frac{F_\alpha}{n}\right)H\left(\frac{F_\alpha/n}{2+F_\alpha/n}\right)+\frac{F_\alpha}{n} \geq \alpha + o(1),$$

where H is the binary entropy function. The expression on the left side of the inequality is strictly increasing in $\frac{F_\alpha}{n}$, and it is less than 0.99 if we substitute $\frac{F_\alpha}{n}$ by 0.045. If we let $\alpha = 0.99$, it follows that for sufficiently large n, we have $\frac{F_\alpha}{n} \geq 0.045$, thereby establishing the lower bound in Theorem 3.

To prove the upper bound in Theorem 3, we prove the following theorem.

Theorem 7. *The limit* $\lim_{n\to\infty} f(n)/n$ *exists and is* $\leq 17/32$.

Proof. Note that for any positive integers n and m, $f(n+m) \leq f(n)+f(m)+2$. Indeed, we can deduplicate separately the first n bits of the sequence and the last m bits, getting a concatenation of two square-free sequences (of total length at most 6). It then suffices to check that each such concatenation can be deduplicated to its root through at most 2 additional deduplication steps. Therefore, the function $g(n) = f(n) + 2$ is subadditive:

$$\begin{aligned} g(n+m) = f(n+m)+2 &\leq f(n)+f(m)+4 \\ &= g(n)+g(m). \end{aligned}$$

By Fekete's Lemma [9], $g(n)/n$ tends to a limit (which is the infimum over n of $g(n)/n$), and it is clear that the limit of $f(n)/n$ is the same as that of $g(n)/n$.

This proof of the existence of $\lim_{n\to\infty} f(n)/n$ provides a simple way to derive an upper bound for the limit by computing $f(n)$ precisely for some small n. In particular, from Figure 1, we find $\lim_{n\to\infty} f(n)/n \leq (f(32)+2)/32 = 17/32$. □

III. Duplication/Hamming Distance

In this section, we provide bounds on $f_\beta(n)$ for $\beta < 1/2$ and $\beta > 1/2$. We first however present some useful definitions. For $0 \leq \beta < 1$, a *β-repeat of length h* in a binary sequence consists of two consecutive blocks in the sequence, each of length h, such that the Hamming distance between them is at most βh. If $\boldsymbol{uvv'w}$ is a binary sequence, and $\boldsymbol{vv'}$ is a β-repeat, then a *β-deduplication* produces $\boldsymbol{uvw}$ or $\boldsymbol{uv'w}$. Note that in this case the set of roots of s is not necessarily unique, but the length of any root is at most 3, even if $\beta = 0$. The next theorem establishes a sharp phase transition in the behavior of $f_\beta(n)$ at $\beta = 1/2$. Its proof relies on Theorem 9, which guarantees the existence of β-repeats under certain conditions. In what follows, for an integer m, we use $[m]$ to denote $\{1,\ldots,m\}$.

Theorem 8. *If* $\beta < 1/2$, *there exists a constant* $c = c(\beta) > 0$ *such that* $f_\beta(n) \geq cn$. *Furthermore, if* $\beta > 1/2$, *for any constant* $C > \left\lceil\frac{2\beta+1}{2\beta-1}\right\rceil^2$ *and sufficiently large* n, $f_\beta(n) \leq C\ln n$.

Proof. The proof for $\beta < 1/2$ is similar to the proof of the lower bound in Theorem 3. In this case however, to make the deduplication process reversible, for every deduplication we need to record whether it is of the form $\boldsymbol{uvv'w} \xrightarrow{dd} \boldsymbol{uvw}$ or of the form $\boldsymbol{uv'vw} \xrightarrow{dd} \boldsymbol{uvw}$, and we must also encode the sequence $\boldsymbol{v'}$. In the tth deduplication step, we have $|\boldsymbol{v}| = |\boldsymbol{v'}| = h_t$. Since $\boldsymbol{v'}$ is in the Hamming sphere of radius βh_t around $\boldsymbol{v}$, there are at most $2^{h_t H(\beta)}$ options for $\boldsymbol{v'}$. Thus

$$6n\sum_{f=1}^{F_\beta}\binom{n+f}{f}\binom{2n+f}{f}\binom{2n+f+2}{f}2^{nH(\beta)}2^{2f} \geq 2^n,$$

where $F_\beta = f_\beta(n)$ and we have used $\sum_t h_t \leq n$. The desired result then follows since $H(\beta) < 1$.

Suppose $\beta > 1/2$. Let $K = \left\lceil\frac{2\beta+1}{2\beta-1}\right\rceil^2$ and $\epsilon = C - K$. Note that $\epsilon > 0$. By appropriately choosing C_1, we can have $f_\beta(i) \leq \left(K+\frac{\epsilon}{2}\right)\ln i + C_1$ for all $i < M$, where M is sufficiently large and in particular $M > K$. Assuming that this holds also for all $i < n$, where $n \geq M$, we show that it holds for $i = n$. From Theorem 9, every binary sequence $\boldsymbol{s}$ of length n has a β-repeat of length $\ell\lfloor n/K\rfloor$ for some $\ell \in \left[\sqrt{K}\right]$, implying

$$\begin{aligned} f_\beta(\boldsymbol{s}) &\leq f_\beta\left(n-\ell\left\lfloor\frac{n}{K}\right\rfloor\right)+1 \\ &\leq \left(K+\frac{\epsilon}{2}\right)\ln\left(n-\left\lfloor\frac{n}{K}\right\rfloor\right)+1+C_1 \\ &\leq \left(K+\frac{\epsilon}{2}\right)\ln n - \frac{(K+\frac{\epsilon}{2})(n-K)}{Kn}+1+C_1 \\ &\leq \left(K+\frac{\epsilon}{2}\right)\ln n + C_1 \leq C\ln n, \end{aligned}$$

where the last two steps hold for sufficiently large n. Hence, $f_\beta(n) \leq C\ln n$. □

Theorem 9. *If* $\beta > \frac{1}{2}$, *then for any integer* $k \geq \frac{2\beta+1}{2\beta-1}$, *any binary sequence of length* n *contains a* β*-repeat of length* $\ell\lfloor n/k^2\rfloor$ *for some* $\ell \in [k]$.

Proof. Let k be a positive integer to be determined later and put $K = k^2$. Furthermore, let $\boldsymbol{s'} = \boldsymbol{s}_1\cdots\boldsymbol{s}_K$ be a partition of the first KB symbols of $\boldsymbol{s}$ into blocks of length $B = \left\lfloor\frac{n}{K}\right\rfloor$. We now consider as a code the $k+1$ binary vectors

$$\boldsymbol{t}_i = \boldsymbol{s}_i\cdots\boldsymbol{s}_{i+K-k-1}, \qquad (1 \leq i \leq k+1),$$

each of length $m = (K-k)B$. By Plotkin's bound [8, p. 41], the minimum Hamming distance of this code is at most $\left(\frac{1}{2}+\frac{1}{2k}\right)m$. Thus there exist $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$ with $i < j$ with Hamming distance at most $\left(\frac{1}{2}+\frac{1}{2k}\right)m$.

Put $h = (j-i)B$ and let $m' = h\lfloor m/h\rfloor$ be the largest integer which is at most m and is divisible by h. Let $\boldsymbol{t}'_i$ and $\boldsymbol{t}'_j$ consist of the first m' bits of $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$, respectively. The Hamming distance between $\boldsymbol{t}'_i$ and $\boldsymbol{t}'_j$ is clearly still at most $\left(\frac{1}{2}+\frac{1}{2k}\right)m$. But $\left(\frac{1}{2}+\frac{1}{2k}\right)m \leq \left(\frac{1}{2}+\frac{1}{k-1}\right)m'$ since

$$\begin{aligned} \left(\frac{1}{2}+\frac{1}{2k}\right)m &= \left(\frac{1}{2}+\frac{1}{2k}\right)\frac{m}{m'}m' \\ &\leq \left(\frac{1}{2}+\frac{1}{2k}\right)\frac{k}{k-1}m' = \left(\frac{1}{2}+\frac{1}{k-1}\right)m', \end{aligned}$$

where we have used the facts that $m = k(k-1)B$ and

$$\frac{m-m'}{B} < \frac{h}{B} \le k,$$

which since B divides m, m', implies $\frac{m-m'}{B} \le k-1$ and, in turn, $m' \ge m-(k-1)B = (k-1)^2 B$.

Split $\boldsymbol{t}'_i$ and $\boldsymbol{t}'_j$ into blocks of length h each: $\boldsymbol{t}'_i = \boldsymbol{z}_1\boldsymbol{z}_2\cdots\boldsymbol{z}_p$, $\boldsymbol{t}'_j = \boldsymbol{z}_2\boldsymbol{z}_3\cdots\boldsymbol{z}_p\boldsymbol{z}_{p+1}$, where $p = m'/h$. The Hamming distance between $\boldsymbol{t}'_i$ and $\boldsymbol{t}'_j$ is the sum of the Hamming distances between $\boldsymbol{z}_q$ and $\boldsymbol{z}_{q+1}$ as q ranges from 1 to p. Thus, by averaging, there exists an index r so that the Hamming distance between $\boldsymbol{z}_r$ and $\boldsymbol{z}_{r+1}$ is at most $\left(\frac{1}{2}+\frac{1}{k-1}\right)h$. Putting $k \ge \frac{2\beta+1}{2\beta-1}$ so that $\frac{1}{2}+\frac{1}{k-1} \le \beta$ ensures that $\boldsymbol{z}_r\boldsymbol{z}_{r+1}$ is β-repeat of length $h = (j-i)B = (j-i)\lfloor n/K \rfloor$. □

Let a β_h-repeat be a repeat of length h with at most $h\beta_h$ mismatches, i.e., the two blocks are at Hamming distance at most $h\beta_h$. In the preceding theorems and their proofs, in principal, we do not need the maximum number of permitted mismatches to be a linear function of the length of the repeat, so we can apply the same techniques to β_h-repeats with nonlinear relationships:

Theorem 10. *Let $\beta_h^a = \frac{1}{2} + \frac{1}{h^a}$, where $0 < a < 1$ is a constant, and let $f_a(n)$ be the smallest number f such that any binary sequence of length n can be deduplicated in f steps by deduplicating β_h^a-repeats. There exist positive constants c_2, c_3 such that*

$$f_a(n) \le c_2 n^{2a/(1+a)} + c_3. \tag{3}$$

Proof. By making appropriate changes to the proof of Theorem 9, one can show that for $k = \lceil 2n^{a/(1+a)} \rceil$, every binary sequence of sufficiently long length n contains a β_h^a-repeat of length $h = \ell\lfloor n/k^2 \rfloor$, for some $\ell \in [k]$. To do so, we need to prove $\left(\frac{1}{2}+\frac{1}{k-1}\right)h \le \beta_h^a h$ for all h of the form $h = \ell\lfloor n/k^2 \rfloor$, $\ell \in [k]$. This holds since with the aforementioned value of k,

$$\beta^a_{\ell\lfloor n/k^2\rfloor} = \frac{1}{2} + \frac{1}{(\ell\lfloor n/k^2\rfloor)^a} \ge \frac{1}{2} + \frac{1}{(k\lfloor n/k^2\rfloor)^a} \ge \frac{1}{2} + \frac{1}{k-1},$$

for all $\ell \in [k]$ and sufficiently large n.

We can now prove (3) by induction. Clearly, for any M, there exist constants c_2, c_3 such that $f_a(i) \le c_2 i^{2a/(1+a)} + c_3$ for all $i \le M$. Choose M to be sufficiently large as to satisfy the requirements of the rest of the proof. Fix $n > M$ and assume that $f_a(i) \le c_2 i^{2a/(1+a)} + c_3$ for all $i < n$. Since in every sequence of length n, there exists a β_h^a-repeat with $h = \ell\lfloor n/k^2 \rfloor$, for some $\ell \in [k]$ and $k = \lceil 2n^{a/(1+a)} \rceil$, it holds that

$$\begin{aligned}
f_a(n) &\le 1 + c_2\big(n - \ell\lfloor n/k^2\rfloor\big)^{2a/(1+a)} + c_3 \\
&\le 1 + c_2\left(n - \frac{1}{5}n^{\frac{1-a}{1+a}}\right)^{2a/(1+a)} + c_3 \\
&= 1 + c_2 n^{2a/(1+a)}\left(1 - \frac{1}{5}n^{-\frac{2a}{1+a}}\right)^{2a/(1+a)} + c_3 \\
&\le 1 + c_2 n^{2a/(1+a)}\left(1 - \frac{2a}{5(1+a)}n^{-\frac{2a}{1+a}}\right) + c_3 \\
&= c_2 n^{2a/(1+a)} + \left(1 - \frac{2ac_2}{5(1+a)}\right) + c_3,
\end{aligned}$$

where the inequalities hold for sufficiently large n. Noting that we can choose c_2 to be arbitrarily large completes the proof. □

IV. Open Problems

We now describe some of the open problems related to extremal values of duplication distance. First, while we have presented bounds on $\lim_n \frac{f(n)}{n}$, its value is unknown. Furthermore, although the lower bound $f(\boldsymbol{s}) \ge 0.045n$ is valid for all but an exponentially small fraction of sequences of length n, we have not been able to find an explicit set of arbitrarily long sequences whose distance to the root is linear in n. A related problem to identifying sequences with large deduplication distance is improving bounds on $f(\boldsymbol{s})$ that depend on the structure of $\boldsymbol{s}$, such as the bound given in Lemma 1, relating $f(\boldsymbol{s})$ to the number of distinct k-mers of $\boldsymbol{s}$.

While we showed at $\beta = 1/2$, $f_\beta(n)$ transitions from a linear dependence on n to a logarithmic one, the behavior at $\beta = 1/2$ is not known. Furthermore, we can alter the setting by decoupling mismatches and repeats, i.e., one sequence is taken to another through substitutions and exact duplications, with limitations on the number of substitutions. We can then study the same problems as the ones in this paper as well as new problems, e.g., the minimum number symbol changes required to generate a sequence with a logarithmic number of duplications.

Acknowledgment

The authors would like to thank anonymous reviewers whose comments improved the presentation of this paper. This work was supported in part by the NSF Expeditions in Computing Program (The Molecular Programming Project), by a USA-Israeli BSF grant 2012/107, by an ISF grant 620/13, and by the Israeli I-Core program.

References

[1] F. Farnoud, M. Schwartz, and J. Bruck, "The capacity of string-duplication systems," in *Proc. IEEE Int. Symp. Information Theory*, Honolulu, HI, USA, Jun. 2014, pp. 1301–1305.

[2] ——, "The capacity of string-duplication systems," *To appear in IEEE Trans. Information Theory*.

[3] S. Jain, F. Farnoud, and J. Bruck, "Capacity and expressiveness of genomic tandem duplication," in *Proc. IEEE Int. Symp. Information Theory*, Hong Kong, China, Jun. 2015.

[4] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[5] G. Benson and L. Dong, "Reconstructing the duplication history of a tandem repeat." in *ISMB*, 1999, pp. 44–53.

[6] O. Gascuel, D. Bertrand, and O. Elemento, *Reconstructing the duplication history of tandemly repeated sequences*, O. Gascuel, Ed. Oxford: Oxford University Press, 2005.

[7] A. Thue, "Über unendliche zeichenreihen," *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl., Christiana*, 1906.

[8] F. J. MacWilliams and N. J. A. Sloane, *The theory of error correcting codes*. New York: Elsevier/North-Holland Inc., 1977.

[9] J. M. Steele, *Probability Theory and Combinatorial Optimization*. Society for Industrial and Applied Mathematics, 1997. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/1.9781611970029