Duplication Distance to the Root for Binary Sequences

Noga Alon, Jehoshua Bruck, *Fellow, IEEE,* Farzad Farnoud, *Member, IEEE,* and Siddharth Jain, *Student Member, IEEE*

Abstract

We study the tandem duplication distance between binary sequences and their roots. In other words, the quantity of interest is the number of tandem duplication operations of the form $\boldsymbol{x} = \boldsymbol{abc} \to \boldsymbol{y} = \boldsymbol{abbc}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are sequences and $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are their substrings, needed to generate a binary sequence of length n starting from a square-free sequence from the set $\{0, 1, 01, 10, 010, 101\}$. This problem is a restricted case of finding the duplication/deduplication distance between two sequences, defined as the minimum number of duplication and deduplication operations required to transform one sequence to the other. We consider both exact and approximate tandem duplications. 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 that the maximum distance has a sharp transition from linear in n to logarithmic at $\beta = 1/2$. We also study the duplication distance to the root for sequences with a given root and for special classes of sequences, namely, the de Bruijn sequences, the Thue-Morse sequence, and the Fibbonaci words. The problem is motivated by genomic tandem duplication mutations and the smallest number of tandem duplication events required to generate a given biological sequence.

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 (e.g., $A\underline{CGT} \to A\underline{CGCGT}$); and *point mutations* or *substitutions*, which substitute one base in the sequence by another (e.g., $A\underline{C}GT \to A\underline{T}GT$). 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 is the study of how genomic sequences are generated through mutations. In this paper, we focus on tandem duplication mutations and tandem repeats, which form about 3% of the human genome [10], and study the minimum number of mutation events that can create a given sequence. More 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.

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

Jehoshua Bruck is with the Electrical Engineering Department, California Institute of Technology, Pasadena, CA, 91125, Email: bruck@caltech.edu.

Farzad Farnoud is with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA, 22903, Email: farzad@virginia.edu. He was with the Electrical Engineering Department, California Institute of Technology.

Siddharth Jain is with the Electrical Engineering Department, California Institute of Technology, Pasadena, CA, 91125, Email: sidjain@caltech.edu.

This paper was presented in part at 2016 IEEE International Symposium on Information Theory in Barcelona, Spain.

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 $12\underline{13413}\underline{4}51$ contains the repeat 134134 of length 3. A repeat of length h may be created through a (*tandem*) *duplication of length* h, e.g., $1213451 \xrightarrow{d} 1213413451$, where $\xrightarrow{d}$ denotes a duplication operation. On the other hand, a repeat may be removed through a (*tandem*) *deduplication of length* h, i.e., by removing one of the two adjacent identical blocks, e.g., $1213413451 \xrightarrow{dd} 1213451$.

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 $\boldsymbol{y}$, denoted $\boldsymbol{x} = \mathrm{root}(\boldsymbol{y})$, 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 sequences to their roots. 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 a sequence present in the genome of an organism.

Our attention here 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, $n \in \mathbb{N}$, are 0 and 1, respectively. For every other binary sequence $\boldsymbol{s}$ of length n, the root of $\boldsymbol{s}$ is the sequence 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

$$101 \xrightarrow{d} \underline{10}10\underline{1} \xrightarrow{d} 1010\underline{11} \xrightarrow{d} 1\underline{00}1011 = \boldsymbol{s}.$$

A *run* in a sequence is a maximal substring consisting of one or more copies of a single symbol. Through duplication, we can generate every binary sequence from its root by first creating the correct number of runs of appropriate symbols. For example, since $\boldsymbol{s} = 1001011$ has 5 runs, the first being a run of the symbol 1, we first generate 10101 through duplication. It is not difficult to see that this is always possible. The next step is then to extend each run so that it has the appropriate length.

In the proofs in the paper, it is often helpful to think of the distance to the root in terms of converting a sequence to its root via a sequence of deduplications, e.g. the sequence $\boldsymbol{s}$ above can be *deduplicated to* its root as

$$\boldsymbol{s} = 1\underline{00}1011 \xrightarrow{dd} 1010\underline{11} \xrightarrow{dd} \underline{10}10\underline{1} \xrightarrow{dd} 101 = \mathrm{root}(\boldsymbol{s}).$$

We note that a celebrated result by Thue from 1906 [17] states that for alphabets of size ≥ 3, there is an infinite square-free sequence. Thus, in contrast to the binary alphabet, the set of roots for such alphabets is infinite since each substring of Thue's sequence is square-free.

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. Table I, 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 $f(n)$. We also consider a variation of the duplication distance, referred to as the *approximate-duplication distance*, where the duplication

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

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$f(n)$	0	1	2	2	3	4	4	5	6	6	7	7	8	8	9	9
n	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
$f(n)$	10	10	11	11	11	12	12	12	13	13	13	14	14	14	15	15

TABLE I
$f(n)$ FOR $1 \leq n \leq 32$.

process is imprecise and so the inserted block is not necessarily an exact copy. Specifically, the *β-approximate-duplication 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. This 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 $f_\beta(n)$. We provide bounds on $f_\beta(n)$ and in particular show that there is a sharp transition in the behavior of f_β at $\beta = 1/2$.

Since each binary sequence has a unique root in the set $\{0, 1, 01, 10, 010, 101\}$, the set of sequences can be partitioned based on their roots. In the paper, we also study the duplication distance to the root for sequences based on the part they belong to, that is, we consider $f_{\boldsymbol{\sigma}}(n)$ for $\boldsymbol{\sigma} \in \{0, 1, 01, 10, 010, 101\}$, where $f_{\boldsymbol{\sigma}}(n) = \max\{f(\boldsymbol{s}) : \mathrm{root}(\boldsymbol{s}) = \sigma, |\boldsymbol{s}| = n\}$.

The rest of the paper is structured as follows. In the next two subsections, we summarize the results of the paper and describe the related work. Then, in Section II, we prove the bounds on $f(n)$ and discuss some variants, as well as special classes of sequences. In Section III, we discuss duplication distance for special class of sequence generating systems called Lindenmayer Systems. In Sections IV and V, we study the approximate-duplication distance to the root and the duplication distance for different roots, respectively. Finally, several open problems and possible future directions are presented in Section VI.

A. Results

In this subsection, we present the main results of the paper. The proofs, unless they are very short, are postponed to later sections.

Suppose the root of $\boldsymbol{s}$ is $\boldsymbol{\sigma} \in \{0, 1, 01, 10, 010, 101\}$. It is easy to see that

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

While the above lower bound is tight in the sense that there exist $\boldsymbol{\sigma}$ and $\boldsymbol{s}$ that satisfy it with equality, e.g., $\boldsymbol{s} = 0^{2^k}$ and $\boldsymbol{\sigma} = 0$, we show there is a positive constant c such that for most sequences of length n, the duplication distance to the root is bounded below by cn. We also improve the upper bound.

Theorem 1. *The limit $\lim_{n\to\infty} f(n)/n$ exists and*

$$0.045 \leq \lim_{n\to\infty} \frac{f(n)}{n} \leq \frac{2}{5}\,.$$

Furthermore, for sufficiently large n, $f(\boldsymbol{s}) \geq 0.045n$ for all but an exponentially small fraction of sequences $\boldsymbol{s}$ of length n; and $f(n) \leq 2n/5 + 15$.

Although the linear lower bound on the duplication distance to the root holds for almost all sequences, finding a specific family of sequences that satisfy it appears to be difficult. The next

lemma and its corollary give the best known construction for a family with large distance to the root, namely, this family achieves distance $\Omega(n/\log n)$.

Lemma 2. *Consider a sequence $\boldsymbol{s}$ and a positive integer $k \geq 4$, and let $K(\boldsymbol{s})$ denote the number of distinct k-mers (sequences of length k) occurring in $\boldsymbol{s}$. We have*

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

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 $k \geq 4$. □

A *binary De Bruijn sequence* [2] of order k is a binary sequence of length $n = 2^k$ that when viewed cyclically contains every possible binary sequence of length k as a substring exactly once. For example, 0011 and 00010111 are De Bruijn sequences of order 2 and order 3, respectively. A binary De Bruijn sequence of order k and length $n = 2^k$ has precisely $n - k + 1$ distinct k-mers. Hence, we have the following corollary.

Corollary 3. *For any binary De Bruijn sequence $\boldsymbol{s}$ of order k (which has length $n = 2^k$), we have*

$$f(\boldsymbol{s}) \geq \frac{n - \log_2 n}{\log_2 n} \ .$$

It is worth noting that using the same technique as the proof of $f(n) = \Omega(n)$ in Theorem 1, and the fact that there are at least $\frac{2^{n/2}}{n}$ De Bruijn sequences of length n when n is a power of two,[2] one can show that the largest duplication distance for De Bruijn sequences grows linearly in their length.

A question arising from observing that $f(n) = \Theta(n)$ is that how does allowing mismatches in the duplication process affect the distance to the root. In particular, for what values of β, is $f_\beta(n)$ linear in n and for what values is it logarithmic? The next theorem establishes that there is a sharp transition at $\beta = 1/2$.

Theorem 4. *If $\beta < 1/2$, then 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,

$$f_\beta(n) \leq C \ln n.$$

Finally, we establish that the limit of $\frac{f(n)}{n}$ is the same if we consider only sequences with root 10 or only sequences with root 101.

Theorem 5. *The limits $\lim_n \frac{f_{10}(n)}{n}$ and $\lim_n \frac{f_{101}(n)}{n}$ exist and are equal to $\lim_n \frac{f(n)}{n}$.*

[2]If De Bruijn seqences are defined cyclically as opposed to linearly, there are exactly $\frac{2^{n/2}}{n}$ De Bruijn sequences of length n

B. Related Work

Tandem duplications and repeats in sequences have been studied from a variety of points of view. One of the most relevant to this work is the study of estimating the tandem duplication history of a given sequence, i.e., a sequence of duplication events that may have generated the sequence, see e.g., [1], [16], [7]. While the afforementioned works study the problem from an algorithmic point of view, in this paper, we are focused on extremal distance values for binary sequences. Furthermore, [16], [7] have a more restrictive duplication model than that of the present paper.

Another aspect, the study of the ability of duplication mutations to generate diversity, has been recently investigated from an information-theoretic point of view [6], [8]. In particular, [6] models sequences generated from a starting "seed" through different types of duplications as sequence 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 [6], [8] 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.

Finally, we mention that the stochastic behavior of certain duplication systems has been studied in [3], [5], and error-correcting codes for combating duplication errors have been introduced in [9].

II. Bounds on $f(n)$

Theorem 1. *The limit* $\lim_{n\to\infty} f(n)/n$ *exists and*

$$0.045 \leq \lim_{n\to\infty} \frac{f(n)}{n} \leq \frac{2}{5}\ .$$

Furthermore, for sufficiently large n, $f(\boldsymbol{s}) \geq 0.045n$ *for all but an exponentially small fraction of sequences* $\boldsymbol{s}$ *of length* n*; and* $f(n) \leq 2n/5 + 15$.

The lower bound of Theorem 1 is proved with the help of Theorem 6, and its upper bound uses Theorem 9. These theorems are stated next.

Theorem 6. *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*

$$6n \sum_{f=1}^{F_\alpha} \binom{n+f}{f} \binom{2n+f}{f} \binom{2n+f+2}{f} 2^f \geq 2^{n\alpha} - 1. \tag{1}$$

Before stating the proof, we present some background, definitions, and a useful claim, 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 $\boldsymbol{s} = \boldsymbol{s}_0 \xrightarrow{dd} \boldsymbol{s}_1 \xrightarrow{dd} \boldsymbol{s}_2 \xrightarrow{dd} \cdots \xrightarrow{dd} \boldsymbol{s}_f = \mathrm{root}(\boldsymbol{s})$, 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 $\boldsymbol{s} = 123\underline{13}\underline{4}13451$,

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

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

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 1.

Definition 7. A deduplication process $\boldsymbol{s} = \boldsymbol{s}_0 \xrightarrow{dd} \boldsymbol{s}_1 \xrightarrow{dd} \boldsymbol{s}_2 \xrightarrow{dd} \cdots \xrightarrow{dd} \boldsymbol{s}_f = \mathrm{root}(\boldsymbol{s})$ 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 $1 \leq t < f$, if $i_{t+1} < i_t$ then $i_{t+1} + 2h_{t+1} \geq i_t$.

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 8. *For any deduplication process $\boldsymbol{s} = \boldsymbol{s}_0 \xrightarrow{dd} \boldsymbol{s}_1 \xrightarrow{dd} \boldsymbol{s}_2 \xrightarrow{dd} \cdots \xrightarrow{dd} \boldsymbol{s}_f = \mathrm{root}(\boldsymbol{s})$ 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 $1 \leq t < q \leq f$, and $i_q < i_t$. We claim that this process is normal. Indeed, otherwise there is some t, $1 \leq t < f$ 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 $\boldsymbol{s}_f = \mathrm{root}(\boldsymbol{s})$ stay the same. This contradicts the minimality in the choice of the process, establishing the claim. □

We now turn to the proof of Theorem 6.

Proof of Theorem 6. 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 8, 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 $f \leq F_\alpha$. 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 $1 \leq t < n$. 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 $i_{t+1} + 2h_{t+1} \geq i_t$, that is $i_t - i_{t+1} \leq 2h_{t+1}$. 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

$$\sum_{t:i_t-i_{t+1}\le 0}(i_t - i_{t+1}) = (i_1 - i_f) - \sum_{t:i_t-i_{t+1}>0}(i_t - i_{t+1}) > -2 - 2n.$$

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 $|U_\alpha| \ge 2^{n\alpha} - 1$, implies the assertion of Theorem 6. □

Since $\binom{p}{q} \le 2^{pH(q/p)}$ for positive integers $0 < q < p$ [12, p. 309], Theorem 6 implies that

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

where H is the binary entropy function, $H(x) = -x\log_2 x - (1-x)\log_2(1-x)$. 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 $\frac{F_\alpha}{n} \ge 0.045$, thereby establishing the lower bound in Theorem 1.

To prove the upper bound in Theorem 1, we prove the following theorem.

Theorem 9. *The limit* $\lim_{n\to\infty} f(n)/n$ *exists and for all* n, $f(n) \le \frac{2}{5}n + 15$.

Proof. Note that for any positive integers n and m, $f(n+m) \le f(n) + f(m) + 2$. Indeed, given a sequences of length $n + m$ we can deduplicate separately its first n bits and its 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:

$$g(n+m) = f(n+m) + 2 \le f(n) + f(m) + 4 = g(n) + g(m).$$

Now, by Fekete's Lemma [15], $g(n)/n$ 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 $g(n)/n$. We term this limit the *binary duplicatoin constant*.

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 Table I, we find $\lim_{n\to\infty} f(n)/n \le (f(32) + 2)/32 = 17/32$. We can improve upon this result as follows.

For positive integers n, m, let $f(n, m)$ be the smallest number k such that every sequence of length n can be converted to a sequences of length at most m via k deduplication steps. A sequence of length n can be converted to its root by first repeatedly converting its a-substrings to substrings of length at most b via $f(a, b)$ deduplication steps. Thus for integers $a > b > 0$, we have

$$f(n) \le \frac{f(a,b)}{a-b}n + \max_{i<a} f(i) \tag{3}$$

With the help of a computer we find the values of $f(n, m)$ for $3 \le m < n \le 32$. An illustration is given in Figure 1. In particular we have $\frac{f(32,12)}{20} = \frac{8}{20} = \frac{2}{5}$ from Figure 1 and $\max_{i<32} f(i) = 15$ from Table I, implying $f(n) \le \frac{2}{5}n + 15$. □

Weaker upper bounds on $f(n)$ can be obtained without resorting to computation in the following ways. First, to deduplicate a sequence to its root, we first can deduplicate each block of t consecutive identical bits to a single bit by $\lceil \log_2 t \rceil$ deduplications and then finish in less than

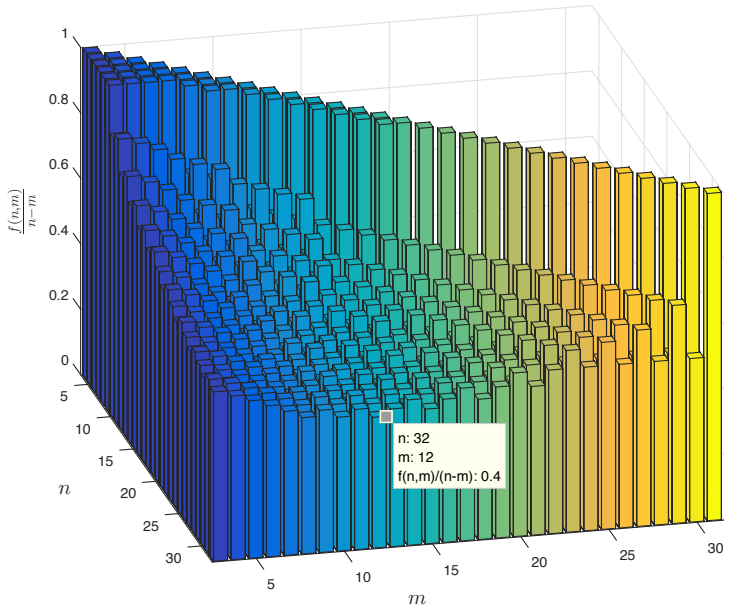


Fig. 1. $\frac{f(n,m)}{n-m}$ for $3 \leq m < n \leq 32$.

$\log_2 n$ additional steps. This shows that for large n , $f(n) \leq \frac{2}{3}n + o(n)$ (the extremal case for this argument is the one in which each block is of size 3). Second, it is known that every binary sequence of length at least 19 contains a repeat of length at least 2 [4], implying that $f(n) \leq \frac{1}{2}n + o(n)$.

Parallel duplication : One can also define the parallel duplication distance to the root by allowing non-overlapping duplications to occur simultanously, with $f'(n)$ being the maxmimum parallel duplication distance to the root of a sequence of length n. Similar to the normal duplication distance it is helpful to think in terms of deduplications. Since each parallel deduplication step decreases the length of a sequence by at most a factor of 2, $f'(n) > \log_2 n - 2$ (and in fact $f'(\boldsymbol{s}) \geq \log_2 n - 2$ for every sequence of length n.) It is not difficult to see that $f'(n) < 2\log_2 n$ by first deduplicating, in parallel, all blocks of identical elements in the sequence to blocks of size 1, and then by deduplicating the resulting alternating sequence to its root.

Partial deduplication: The definition of $f(n, m)$ gives rise to the following question: For a fixed $0 < \alpha \leq 1$, what is $\lim_n \frac{f(n,\lfloor \alpha n \rfloor)}{1-\alpha}$, if it exists? At first glance, one may expect $\lim_n \frac{f(n,\lfloor \alpha n \rfloor)}{1-\alpha}$ to be decreasing in α since if α is large, one may think it is easier to find enough long repeats to reduce the length of the sequence quickly by a factor of $1 - \alpha$. However, we show that $\lim_n \frac{f(n,\lfloor \alpha n \rfloor)}{n(1-\alpha)} = \lim_n \frac{f(n)}{n}$.

Let $\gamma = \lim_n \frac{f(n)}{n}$. For $\epsilon > 0$, there exists k such that for all $n > k$, $f(n) \leq (\gamma + \epsilon)n$. Thus

$$f(n, \lfloor \alpha n \rfloor) \leq f(n - \lfloor \alpha n \rfloor + 3) \leq (\gamma + \epsilon)((1 - \alpha)n + 4). \tag{4}$$

On the other hand, let $\delta = \liminf_n \frac{f(n,\lfloor \alpha n \rfloor)}{(1-\alpha)n}$. For $\epsilon > 0$, there exists k such $f(k, \lfloor \alpha k \rfloor) \leq (\delta + \epsilon)(1 - \alpha)k$. Hence,

$$f(n) \leq \frac{f(k, \lfloor \alpha k \rfloor)}{k - \lfloor \alpha k \rfloor} n + k \leq (\delta + \epsilon)n + k. \tag{5}$$

The result follows by dividing (4) by $(1-\alpha)n$ and taking a $\limsup_n$ and by dividing (5) by n and taking a $\lim_n$.

III. Duplication Distance for L-systems

L-systems, or Lindenmayer systems are sequence rewriting systems developed by Lindemayer in 1968 [11]. He used them in the context of biology to model the growth process of plant development. He introduced context-free as well as context-sensitive L-systems. Here we will discuss distance to the root for sequences arising in context-free L-systems, also known as 0L-systems. A 0L-system comprises three components:

- Alphabet (Σ): An alphabet of symbols used to construct sequences.
- Axiom sequence or initiator ($\boldsymbol{\omega}$): The starting sequence from which a 0L-system is constructed.
- Production rule (h): A rule that constructs new sequences by expanding each symbol in a given sequence into a sequence of symbols. The production rule is represented by the function $h : \Sigma^* \to \Sigma^*$, which for any two sequences $\boldsymbol{a}$ and $\boldsymbol{b} \in \Sigma^*$ satisfies

  $$h(\boldsymbol{ab}) = h(\boldsymbol{a})h(\boldsymbol{b})$$

  where $h(\boldsymbol{a})h(\boldsymbol{b})$ represents the concatenation of $h(\boldsymbol{a})$ and $h(\boldsymbol{b})$. The production rule h can be deterministic or stochastic. Here we consider only deterministic rules. Such 0L-systems with deterministic h are denoted as D0L-systems [13].

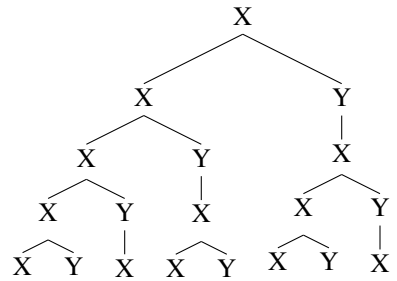
Example 10 (*Fibonacci words*)**.** Consider $\Sigma = \{X, Y\}$, $\boldsymbol{\omega} = X$, and

$$h(X) = XY, \quad h(Y) = X.$$

For this D0L-system, the first 5 sequences are as follows:

$$\begin{aligned}
h^0(\boldsymbol{\omega}) &= X \\
h^1(\boldsymbol{\omega}) &= XY \\
h^2(\boldsymbol{\omega}) &= XYX \\
h^3(\boldsymbol{\omega}) &= XYXXY \\
h^4(\boldsymbol{\omega}) &= XYXXYXYX \\
h^5(\boldsymbol{\omega}) &= XYXXYXYXXYXXY
\end{aligned}$$

This can also be represented by the following tree:



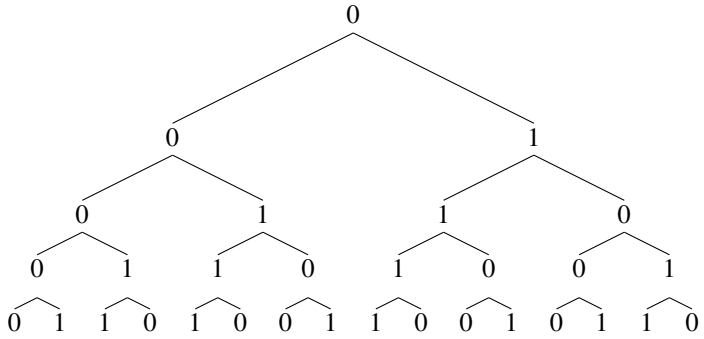
These sequences are called Fibonacci words as they satisfy

$$h^n(\boldsymbol{\omega}) = h^{n-1}(\boldsymbol{\omega})h^{n-2}(\boldsymbol{\omega}) \ \forall \ n \geq 2.$$

Example 11 (*Thue-Morse Sequence*)**.** Let $\Sigma = \{0, 1\}$, $\boldsymbol{\omega} = 0$, and

$$h(0) = 01, \quad h(1) = 10.$$

For this D0L-system the tree of sequence generation is given below:



The sequence generated by this D0L-system are called Thue-Morse sequences. Alternatively, the Thue-Morse sequences can be defined recursively by starting with $\boldsymbol{t}_0 = 0$ and forming $\boldsymbol{t}_{i+1}$ by concatenating $\boldsymbol{t}_i$ and its complement $\overline{\boldsymbol{t}_i}$.

We show that binary D0L-systems, which have production rules of the form $h(0) = \boldsymbol{u}$ and $h(1) = \boldsymbol{v}$, with $\boldsymbol{u}, \boldsymbol{v} \in \{0, 1\}^*$ have a logarithmic distance to their roots.

Lemma 12. *For any binary D0L-system with initiator $\boldsymbol{\omega}$ and production rule h, we have*

$$f(h^r(\boldsymbol{\omega})) = \Theta(\log_2|h^r(\boldsymbol{\omega})|), \qquad \text{as } r \to \infty.$$

Proof. For any sequence $\boldsymbol{t}$, since $f(\boldsymbol{t}) \geq \log_2 |\boldsymbol{t}|$, we have $f(h^r(\boldsymbol{\omega})) \geq \log_2|h^r(\boldsymbol{\omega})|$. It remains to show that $f(h^r(\boldsymbol{\omega})) = O(\log_2|h^r(\boldsymbol{\omega})|)$. We start by proving the following claim.

Claim. *For any binary D0L-system with initiator ω and production rule h, we have*

$$f(h^r(\boldsymbol{\omega})) \leq f\big(h^{r-1}(\boldsymbol{\omega})\big) + c \leq f(\boldsymbol{\omega}) + rc, \tag{6}$$

where $c = \max_{\boldsymbol{z} \in \{0,1,01,10,010,101\}} f(h(\boldsymbol{z}))$.

To prove the claim, let $\boldsymbol{x} = h^{r-1}(\boldsymbol{\omega})$ and $\boldsymbol{y} = h^r(\boldsymbol{\omega})$ and consider the sequence of deduplications that turns x into its root $\boldsymbol{z} \in \{0, 1, 01, 10, 010, 101\}$. We can deduplicate $\boldsymbol{y}$ in a similar manner to $h(\boldsymbol{z})$: For each step in the deduplication process of $\boldsymbol{x}$ that deduplicates a substring $a_1 \cdots a_k a_1 \cdots a_k$ to $a_1 \cdots a_k$, we deduplicate $h(a_1) \cdots h(a_k)h(a_1) \cdots h(a_k)$ to $h(a_1) \cdots h(a_k)$ in the deduplication process of $\boldsymbol{y}$, resulting eventually in $h(\boldsymbol{z})$. This completes the proof of the claim.

We now turn to proving $f(h^r(\boldsymbol{\omega})) = O(\log_2|h^r(\boldsymbol{\omega})|)$. If $|h^r(\boldsymbol{\omega})| = O(1)$, then $f(h^r(\boldsymbol{\omega})) = O(1)$ as well, and there is nothing to prove. If $|h^r(\boldsymbol{\omega})| = 2^{\Omega(r)}$, then $r = O(\log_2|h^r(\boldsymbol{\omega})|)$ and the desired result follows from (6). The last case that we need to consider is when $|h^r(\boldsymbol{\omega})| \to \infty$ but $|h^r(\boldsymbol{\omega})| = 2^{o(r)}$. Without loss of generality, assume $|h(1)| \geq |h(0)|$. Then the condition $|h^r(\boldsymbol{\omega})| = 2^{o(r)}$ can be shown to occur only if the initiator $\boldsymbol{\omega}$ contains at least one occurrence of 1, $h(0) = 0$, and $h(1)$ has exactly one occurrence of 1 and one or more 0s. In this case, the number of 1s in $h^r(\boldsymbol{\omega})$ is constant and again $f(h^r(\boldsymbol{\omega})) = O(\log_2|h^r(\boldsymbol{\omega})|)$. □

The previous lemma shows that the duplication distances to the root for both of Fibonacci words and Thue-Morse sequences are logarithmic in sequence length. This is particularly interesting in the case of the Thue-Morse sequence. Despite the fact that the Thue-Morse sequence grows by

taking the complement, it contains enough repeats to allow a logarithmic distance. Note also that the Thue-Morse sequence is used to generate ternary square-free sequences.

In the next lemma, we give better bounds than those that can be obtained from Lemma 12 or (6) for Thue-Morse and Fibonacci sequences.

Lemma 13. *Let $\boldsymbol{t}_r$ and $\boldsymbol{u}_r$ denote the rth Thue-Morse and Fibonacci words, respectively. For $r \geq 2$, we have*

$$f(\boldsymbol{t}_r) \leq 2r,$$
$$f(\boldsymbol{u}_r) \leq r.$$

Proof. We first prove the upper bound for $\boldsymbol{t}_r$. For $r \geq 3$, we have

$$\begin{aligned} f(\boldsymbol{t}_r) &= f\big(\boldsymbol{t}_{r-1}\overline{\boldsymbol{t}}_{r-1}\big) \\ &= f\big(\boldsymbol{t}_{r-2}\overline{\boldsymbol{t}}_{r-2}\overline{\boldsymbol{t}}_{r-2}\boldsymbol{t}_{r-2}\big) \\ &\leq 1 + f\big(\boldsymbol{t}_{r-2}\overline{\boldsymbol{t}}_{r-2}\boldsymbol{t}_{r-2}\big) \\ &= 1 + f\big(\boldsymbol{t}_{r-3}\overline{\boldsymbol{t}}_{r-3}\overline{\boldsymbol{t}}_{r-3}\boldsymbol{t}_{r-3}\boldsymbol{t}_{r-3}\overline{\boldsymbol{t}}_{r-3}\big) \\ &\leq 3 + f\big(\boldsymbol{t}_{r-3}\overline{\boldsymbol{t}}_{r-3}\boldsymbol{t}_{r-3}\overline{\boldsymbol{t}}_{r-3}\big) \\ &\leq 4 + f\big(\boldsymbol{t}_{r-3}\overline{\boldsymbol{t}}_{r-3}\big) \\ &= 4 + f(\boldsymbol{t}_{r-2}). \end{aligned}$$

If $r \geq 3$ is even, then $f(\boldsymbol{t}_r) \leq 4\frac{r-2}{2} + f(\boldsymbol{t}_2) = 2(r-2)+1 = 2r-3$; and if $r \geq 3$ is odd, then $f(\boldsymbol{t}_r) \leq 4\frac{r-1}{2} + f(\boldsymbol{t}_1) = 2(r-1)$. This completes the proof of the first claim.

We now turn to $f(\boldsymbol{u}_r)$. The rth Fibonacci word can be obtained via the following recursion: $\boldsymbol{u}_r = \boldsymbol{u}_{r-1}\boldsymbol{u}_{r-2}$ for $r \geq 2$ and $\boldsymbol{u}_0 = 0$, $\boldsymbol{u}_1 = 01$. If $r \geq 5$, then

$$\begin{aligned} \boldsymbol{u}_r &= \boldsymbol{u}_{r-1}\boldsymbol{u}_{r-2} \\ &= \boldsymbol{u}_{r-2}\boldsymbol{u}_{r-3}\boldsymbol{u}_{r-3}\boldsymbol{u}_{r-4} \\ &= \boldsymbol{u}_{r-2}\boldsymbol{u}_{r-3}\boldsymbol{u}_{r-4}\boldsymbol{u}_{r-5}\boldsymbol{u}_{r-4} \\ &= \boldsymbol{u}_{r-2}^2\boldsymbol{u}_{r-5}\boldsymbol{u}_{r-4}. \end{aligned}$$

Hence, $f(\boldsymbol{u}_r) \leq 1 + f(\boldsymbol{u}_{r-2}\boldsymbol{u}_{r-5}\boldsymbol{u}_{r-4})$. Noting that $\boldsymbol{u}_{r-2}\boldsymbol{u}_{r-5}\boldsymbol{u}_{r-4} = \boldsymbol{u}_{r-3}\boldsymbol{u}_{r-4}\boldsymbol{u}_{r-5}\boldsymbol{u}_{r-4} = \boldsymbol{u}_{r-3}^2\boldsymbol{u}_{r-4}$, we write

$$\begin{aligned} f(\boldsymbol{u}_r) &\leq 1 + f(\boldsymbol{u}_{r-2}\boldsymbol{u}_{r-5}\boldsymbol{u}_{r-4}) \\ &= 1 + f\big(\boldsymbol{u}_{r-3}^2\boldsymbol{u}_{r-4}\big) \\ &\leq 2 + f(\boldsymbol{u}_{r-3}\boldsymbol{u}_{r-4}) \\ &= 2 + f(\boldsymbol{u}_{r-2}). \end{aligned}$$

Now, if $r \geq 5$ is even, then $f(\boldsymbol{u}_r) \leq (r-4) + f(\boldsymbol{u}_4) \leq r-2$ since $f(\boldsymbol{u}_4) = f(01001010) \leq 2$; and if $r \geq 5$ is odd, then $f(\boldsymbol{u}_r) \leq (r-3) + f(\boldsymbol{u}_3) \leq r-1$ as $f(\boldsymbol{u}_3) = f(01001) \leq 2$. □

IV. Approximate-duplication distance

Recall that $f_\beta(n)$ is the least k such that every sequence of length n can be converted to a square-free sequence in k approxmiate deduplication steps, with at most a β fraction of mismatches in each step. 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 $\boldsymbol{s}$ 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 14, 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 4. *If $\beta < 1/2$, then 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 1. 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 $\boldsymbol{v'}$ [14, Lemma 4.7]. 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 14, 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

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

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

Theorem 14. *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 $B = \lfloor \frac{n}{K} \rfloor$. We now consider as a code [12] 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 [12, 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 $\left(\frac{1}{2}+\frac{1}{2k}\right)m \le \left(\frac{1}{2}+\frac{1}{k-1}\right)m'$ since

$$\left(\frac{1}{2}+\frac{1}{2k}\right)m = \left(\frac{1}{2}+\frac{1}{2k}\right)\frac{m}{m'}m' \overset{(*)}{\le} \left(\frac{1}{2}+\frac{1}{2k}\right)\frac{k}{k-1}m' = \left(\frac{1}{2}+\frac{1}{k-1}\right)m',$$

where $(*)$ can be proved as follows. By the definition of m', we have $m-m' < h$. Additionally, $h \le kB$ since $1 \le i < j \le k+1$. So,

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

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

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 15. *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 to a root in f steps by deduplicating β_h^a-repeats. There exist positive constants c_2, c_3 such that*

$$f_a(n) \le c_2 n^{2a/(1+a)} + c_3. \tag{7}$$

Proof. By making appropriate changes to the proof of Theorem 14, 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 (7) 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

$$f_a(n) \le 1 + c_2\big(n - \ell\lfloor n/k^2 \rfloor\big)^{2a/(1+a)} + c_3$$

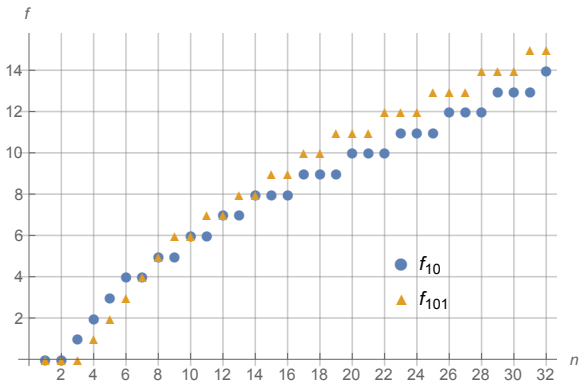


Fig. 2. $f_{10}(n)$ and $f_{101}(n)$ for $1 \le n \le 32$.

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

where the inequalities hold for sufficiently large n. The third inequality follows from Bernoulli's inequality and the the last one follows from the fact that we can choose c_2 to be arbitrarily large. □

V. Duplication distances for different roots

In this section, we study $f_{\boldsymbol{\sigma}}$ for $\boldsymbol{\sigma} \in \{0, 1, 01, 10, 010, 101\}$. It is easy to see that $f_0(n) = f_1(n) = \lceil \log_2 n \rceil$. Clearly $f_{10} = f_{01}$ and $f_{101} = f_{010}$. So we limit our attention to roots $\boldsymbol{\sigma} = 10$ and $\boldsymbol{\sigma} = 101$. Plots for $f_{10}(n)$ and $f_{101}(n)$, obtained through computer search, are given in Figure 2.

Theorem 5. *The limits* $\lim_n \frac{f_{10}(n)}{n}$ *and* $\lim_n \frac{f_{101}(n)}{n}$ *exist and are equal to* $\lim_n \frac{f(n)}{n}$.

Proof. The general approach in this proof is similar to that of the proof of Fekete's lemma in [15]. We prove the theorem for $\lim_n \frac{f_{10}(n)}{n}$. The proof for $\frac{f_{101}(n)}{n}$ is similar.

Let $\gamma = \liminf_n \frac{f_{10}(n)}{n}$ and let $k \ge 3$ be such that $f_{10}(k) + 5 + 2\log_2 k \le k(\gamma + \epsilon)$ for $\epsilon > 0$. Let $\boldsymbol{s}$ be a sequence of length n. Starting from the beginning of $\boldsymbol{s}$, partition it into substrings that are the shortest possible while having length at least k and different symbols at the beginning and the end (so that their root is either 10 or 01). Name these substrings $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{m+1}$, where $|\boldsymbol{s}_i| \ge k$ for $i \le m$ and $1 \le |\boldsymbol{s}_{m+1}| \le k$. Let $s_{i,j}$ denote the jth element of $\boldsymbol{s}_i$. We deduplicate $\boldsymbol{s}$ to its root by first deduplicating its substrings $\boldsymbol{s}_i$ to their roots.

For each substring $\boldsymbol{s}_i$ of the partition, except the last one, we consider the following cases and deduplicate $\boldsymbol{s}_i$ as indicated, where without loss of generality we assume $\boldsymbol{s}_i$ starts with 1 and ends with 0:

- $|\boldsymbol{s}_i| = k$: Deduplicate this substring to 10 in $f_{10}(k)$ steps.
- $|\boldsymbol{s}_i| > k$ and $s_{i,k-1} = 1$: In this case, $\boldsymbol{s}_i = 1\boldsymbol{x}11{,}1^*0$, where $\boldsymbol{x} \in \{0,1\}^{k-3}$, for clarity a comma is placed after the kth element of $\boldsymbol{s}_i$, and a^* denotes that the symbol a appears 0 or more times. We reduce the length of the last run of 1s in $\boldsymbol{s}_i$ by $|\boldsymbol{s}_i|-k$ in $\lceil \log_2(|\boldsymbol{s}_i| - k + 1)\rceil$ deduplication steps to obtain $1\boldsymbol{x}10$. Then deduplicate the result to 10 in $f_{10}(k)$ steps.
- $|\boldsymbol{s}_i| > k$ and $s_{i,k-1} = 0$: In this case, $\boldsymbol{s}_i = 1\boldsymbol{x}01{,}1^*0$, where $\boldsymbol{x} \in \{0,1\}^{k-3}$ and where a comma is placed after the kth element of $\boldsymbol{s}_i$. We reduce the length of the last run of 1s in $\boldsymbol{s}_i$ by $|\boldsymbol{s}_i| - k - 1$ in $\lceil \log_2(|\boldsymbol{s}_i| - k)\rceil$ deduplication steps to obtain $\hat{\boldsymbol{s}}_i = 1\boldsymbol{x}01, 0$ and note that $\hat{\boldsymbol{s}}_i$ has length $k+1$ and ends with 010. Now either $\hat{\boldsymbol{s}}_i$ has a run of length at least 2 or not. If it does, we reduce the length of this run by 1 to obtain a sequence of length k, which we then convert to 10 in $f_{10}(k)$ deduplication steps. If not, then $\hat{\boldsymbol{s}}_i$ is an alternating sequence of the form $101010\cdots10$ which can be deduplicated to 10 in no more than $\lceil \log_2 \frac{k+1}{2}\rceil$ steps.

The resulting sequences has length at most $2m + k$ and can be deduplicated to its root in at most as many steps. We thus have

$$\begin{aligned} f(n) &\le m f_{10}(k) + \sum_{i=1}^{m} \lceil \log_2(|\boldsymbol{s}_i| - k + 1)\rceil + m\left\lceil \log_2 \frac{k+1}{2}\right\rceil + 3m + k \\ &\le m f_{10}(k) + \sum_{i=1}^{m} \log_2|\boldsymbol{s}_i| + m\log_2 k + 5m + k \\ &\le \frac{n}{k} f_{10}(k) + \frac{2n}{k}\log_2 k + 5\frac{n}{k} + k, \end{aligned}$$

where for the last step we have used the fact that

$$\sum_{i=1}^{m} \log_2|\boldsymbol{s}_i| \le m\log_2(n/m) \le \frac{n}{k}\log_2 k$$

which holds since $\sum_{i=1}^{m}|\boldsymbol{s}_i| \le n$, $\frac{d}{dm} m\log_2 \frac{n}{m} > 0$ and $m \le \frac{n}{k}$. It follows that

$$\frac{f(n)}{n} \le \frac{f_{10}(k)}{k} + \frac{2\log_2 k}{k} + \frac{5}{k} + \frac{k}{n} \le \gamma + \epsilon + \frac{k}{n}\ .$$

Taking $\lim$ of both sides and noting that $\epsilon > 0$ is arbitrary proves that $\lim_n \frac{f(n)}{n} \le \liminf_n \frac{f_{10}(n)}{n}$. On the other hand, it is clear that $\limsup_n \frac{f_{10}(n)}{n} \le \lim_n \frac{f(n)}{n}$. Hence, $\lim_n \frac{f(n)}{n} = \lim_n \frac{f_{10}(n)}{n}$. Similar arguments hold for $f_{101}(n)$. □

VI. Open Problems

We now describe some of the open problems related to extremal values of duplication distance to the root. First, the binary duplication constant, $\lim_n \frac{f(n)}{n}$ is unknown. It is also interesting to find bounds tighter than the one given in Theorem 1, namely $0.045 \le \lim \frac{f(n)}{n} \le 0.4$. 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 family of sequences whose distance is linear in n. A related problem to identifying sequences with large duplication distance is improving bounds on $f(\boldsymbol{s})$ that depend on the structure of $\boldsymbol{s}$, such as the bound given in Lemma 2, relating $f(\boldsymbol{s})$ to the number of distinct k-mers of $\boldsymbol{s}$.

While we showed in our study of approximate duplication that 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 duplications and substitutions, i.e., we generate the sequence through exact duplications and substitutions, possibly with limitations on the number

of substitutions. We can then study the same problems as the ones we have in this paper as well as new problems, e.g., the minimum number substitutions required to generate the sequence via a logarithmic number of duplication steps.

A different strand of problems are algorithmic in nature, including designing an algorithm that can efficiently find or approximate the duplication distance to the root and provide a duplication process of the appropriate length. The computational complexity of these tasks is also not known. Similar questions may be asked for approximate duplication, or duplication along with substitution. These problems are important because of their potential application in determining the sequence of duplications and point mutations that may have resulted in a particular segment of an organism's DNA.

ACKNOWLEDGMENT

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] G. BENSON AND L. DONG, *Reconstructing the duplication history of a tandem repeat.*, in ISMB, 1999, pp. 44–53.
[2] N. G. DE BRUIJN, *A combinatorial problem*, Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam, 49 (1946), pp. 758–764. Available: http://repository.tue.nl/415282b7-6c10-4b9f-9624-4437629cc621.
[3] O. ELISHCO, F. FARNOUD, M. SCHWARTZ, AND J. BRUCK, *The capacity of some Pólya string models*, in Proc. IEEE Int. Symp. Information Theory (ISIT), Barcelona, Spain, July 2016, pp. 270–274.
[4] R. C. ENTRINGER, D. E. JACKSON, AND J. SCHATZ, *On nonrepetitive sequences*, J. Combinatorial Theory, Series A, 16 (1974), pp. 159–164.
[5] F. FARNOUD, M. SCHWARTZ, AND J. BRUCK, *A stochastic model for genomic interspersed duplication*, in Proc. IEEE Int. Symp. Information Theory, Hong Kong, China, June 2015, pp. 904–908.
[6] F. FARNOUD, M. SCHWARTZ, AND J. BRUCK, *The capacity of string-duplication systems*, IEEE Trans. Information Theory, 62 (2016), pp. 811–824 (conference version appeared in Proc. of IEEE Int. Symp. on Information Theory (ISIT), Honolulu, HI, June–July 2014).
[7] O. GASCUEL, D. BERTRAND, AND O. ELEMENTO, *Mathematics of Evolution and Phylogeny*, Oxford University Press, Oxford, 2005.
[8] S. JAIN, F. FARNOUD, AND J. BRUCK, *Capacity and expressiveness of genomic tandem duplication*, in Proc. IEEE Int. Symp. Information Theory, Hong Kong, China, June 2015.
[9] S. JAIN, F. FARNOUD, M. SCHWARTZ, AND J. BRUCK, *Duplication-correcting codes for data storage in the dna of living organisms*, in Proc. IEEE Int. Symp. Information Theory (ISIT), Barcelona, Spain, July 2016, pp. 1028–1032.
[10] 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, 409 (2001), pp. 860–921.
[11] A. LINDENMAYER, *Mathematical models for cellular interactions in development*, Theoretical Biology, 18 (1968), pp. 300–315.
[12] F. J. MACWILLIAMS AND N. J. A. SLOANE, *The theory of error correcting codes*, Elsevier/North-Holland Inc., New York, 1977.
[13] P. PRUSINKIEWICZ AND A. LINDENMAYER, *The algorithmic beauty of plants*, Springer–Verlag, 1990.
[14] R. ROTH, *Introduction to coding theory*, Cambridge University Press, 2006.
[15] J. M. STEELE, *Probability Theory and Combinatorial Optimization*, Society for Industrial and Applied Mathematics, 1997.
[16] M. TANG, M. WATERMAN, AND S. YOOSEPH, *Zinc finger gene clusters and tandem gene duplication*, Journal of Computational Biology, 9 (2002), pp. 429–446.
[17] A. THUE, *Über unendliche zeichenreihen*, Norske Vid. Selsk. Skr. I. Mat. Nat. Kl., Christiana, (1906).