The Capacity of String-Replication Systems

Farzad Farnoud (Hassanzadeh)
Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, U.S.A.
farnoud@caltech.edu

Moshe Schwartz
Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer Sheva 8410501, Israel
schwartz@ee.bgu.ac.il

Jehoshua Bruck
Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, U.S.A.
bruck@paradise.caltech.edu

***Abstract*—It is known that the majority of the human genome consists of repeated sequences. Furthermore, it is believed that a significant part of the rest of the genome also originated from repeated sequences and has mutated to its current form. In this paper, we investigate the possibility of constructing an exponentially large number of sequences from a short initial sequence and simple replication rules, including those resembling genomic replication processes. In other words, our goal is to find out the capacity, or the expressive power, of these string-replication systems. Our results include exact capacities, and bounds on the capacities, of four fundamental string-replication systems.**

I. Introduction

More than 50% of the human genome consists of *repeated sequences* [5]. An important class of these repeated sequences are *interspersed repeats*, which are caused by *transposons*. A transposon, or a "jumping gene", is a segment of DNA that can "copy and paste" or "cut and paste" itself into new positions of the genome. Currently, 45% of the human genome is known to consist of transposon-driven repeats [5].

A second type of repeats are *tandem repeats*, generally thought to be caused by *slipped-strand mispairings* [10]. A slipped-strand mispairing is said to occur when, during DNA synthesis, one strand in a DNA duplex becomes misaligned with the other. These mispairings may lead to deletions or insertion of a repeated sequence [8]. While tandem repeats are known to constitute only 3% of the human genome, they cause important phenomena such as chromosome fragility, expansion diseases, silencing genes [11], and rapid morphological variation [3].

While interspersed repeats and random repeats together account for a significant part of the human genome, it is likely that a substantial portion of the unique genome, the part that is not known to contain repeated sequences, also has its origins in ancient repeated sequences that are no longer recognizable due to change over time [5], [11].

Motivated by the prevalence and the significance of repeated sequences and the fact that much of our unique DNA was likely originally repeated sequences, in this paper we study the *capacity* of *string-replication systems* with simple replication rules including rules that resemble the repeat-producing genomic processes, namely duplication of transposons and duplication caused by slipped-strand mispairings. A string-replication system, to be defined formally later, consists of a set of rewriting rules, an initial sequence, and all sequences that can be obtained by a applying the rules to the initial sequence a finite number of times. The notion of capacity, defined later in the paper, represents the average number of bits per symbol that can asymptotically be encoded by the sequences in a string-replication system, and thus illustrates the expressive power and the diversity of that system.

In this paper, we consider four replication rules. The first is the *end replication* rule, which allows substrings of a certain length k to be appended to the end of previous sequences. For example, if $k = 3$ we may construct the sequence TCATGCCAT from TCATGC. While this rule is not biologically motivated, we present it first because of the simplicity of proving the related results. In particular, we show that nearly all sequences with the same alphabet as the initial sequence can be generated with this rule.

The second rule is called *tandem replication* and allows a substring of length k to be replicated next to its original position. For example, for $k = 3$, from the sequence TCATGC, one can generate TCATCATGC. We show that this rule has capacity zero regardless of the initial sequence. However, if one allows substrings of all length larger than a given value to be copied, the capacity becomes positive except in trivial cases.

The third rule is *reversed tandem replication*, which is similar to tandem replication except that the copy is reversed before insertion. For example, in our previous example, the sequence TCATTACGC can be generated. Here, the capacity is zero only in the trivial case in which the initial sequence consists of only one unique symbol.

The last rule is *replication with a gap*, where the copy of a substring of a given length k can be inserted after k' symbols. This rule is motivated by the fact that transposons may insert themselves in places far from their original positions. As an example, for $k = 3$ and $k' = 1$, from TCATGC, one can obtain TCATGCATC. For this rule, we show that the capacity is zero if and only if the initial sequence is periodic with period equal to the greatest common divisor of k and k'.

We note that tandem replication has been already studied in a series of papers [1], [2], [6], [7]. However, this was done in the context of the theory of formal languages, and the goal of these studies was mainly to determine their place in the

This work was supported in part by the NSF Expeditions in Computing Program (The Molecular Programming Project).

Chomsky hierarchy of formal languages.

In the next section, we present the preliminaries and in the following four sections, we present the results for each of the aforementioned replication rules.

II. Preliminaries

Let Σ be some finite alphabet. We recall some useful notation commonly used in the theory of formal languages. An n-string $x = x_1x_2\dots x_n \in \Sigma^n$ is a finite sequence of alphabet symbols, $x_i \in \Sigma$. We say n is the length of x and denote it by $|x| = n$. For two strings, $x \in \Sigma^n$ and $y \in \Sigma^m$, their concatenation is denoted by $xy \in \Sigma^{n+m}$. The set of all finite strings over the alphabet Σ is denoted by Σ^*. We say $v \in \Sigma^*$ is a *substring* of x if $x = uvw$, where $u, w \in \Sigma^*$. The *alpha-representation* of a string s, denoted by $R(s)$, is the set of all letters from Σ making up s. Thus, $R(s) \subseteq \Sigma$. The *alpha-diversity* of s is the size of the alpha-representation of s, denoted by $\delta(s) = |R(s)|$. Furthermore, let the number of occurrences of a symbol $a \in \Sigma$ in a sequence $s \in \Sigma^*$ be denoted by $n_x(a)$. The unique empty word of length 0 is denoted by ϵ.

Given a set $S \subseteq \Sigma^*$, we denote

$$S^* = \{w_1w_2\dots w_m \mid w_i \in S, m \geqslant 0\},$$

whereas

$$S^+ = \{w_1w_2\dots w_m \mid w_i \in S, m \geqslant 1\}.$$

For any $x \in \Sigma^*$, $|x| = n \geqslant m$, the m-suffix of x is $w \in \Sigma^m$, such that $x = vw$ for some $v \in \Sigma^*$. Similarly, the m-prefix of x is $u \in \Sigma^m$, where $x = uv$ for some $u \in \Sigma^*$.

A *string system* S is a subset $S \subseteq \Sigma^*$. For any integer n, we denote by $N_S(n)$ the set of length n strings in S, i.e.,

$$N_S(n) = |S \cap \Sigma^n|.$$

The *capacity* of a string system S is defined by

$$\mathsf{cap}(S) = \limsup_{n\to\infty} \frac{\log_2 N_S(n)}{n}.$$

A *string-replication system* is a tuple $S = (\Sigma, s, \mathcal{T})$, where Σ is a finite alphabet, $s \in \Sigma^*$ is a finite string (which we will use to start the replication process), and where $\mathcal{T}$ is a set of functions such that each $T \in \mathcal{T}$ is a mapping from Σ^* to Σ^* that defines a string-replication rule. The resulting string system S, induced by $(\Sigma, s, \mathcal{T})$, is defined as the closure of the string-replication functions $\mathcal{T}$ on the initial string set $\{s\}$, i.e., S is the minimal set for which $s \in S$, and for each $s' \in S$ and $T \in \mathcal{T}$ we also have $T(s') \in S$.

III. End Replication

We define the end-replication function, $T^{\text{end}}_{i,k} : \Sigma^* \to \Sigma^*$, as follows:

$$T^{\text{end}}_{i,k}(x) = \begin{cases} uvwv & \text{if } x = uvw,\ |u| = i,\ |v| = k \\ x & \text{otherwise.} \end{cases}$$

We also define two sets of these functions which will be used later:

$$\mathcal{T}^{\text{end}}_k = \left\{ T^{\text{end}}_{i,k} \mid i \geqslant 0 \right\}$$
$$\mathcal{T}^{\text{end}}_{\geqslant k} = \left\{ T^{\text{end}}_{i,k'} \mid i \geqslant 0, k' \geqslant k \right\}$$

Intuitively, in the end-replication system, the transformations replicate a substring of length k and append the replicated substring to the end of the original string.

Theorem 1. *Let Σ be any finite alphabet, $k \geqslant 1$ any integer, and $s \in \Sigma^*$, $|s| \geqslant k$. Then for $S^{\text{end}}_k = (\Sigma, s, \mathcal{T}^{\text{end}}_k)$,*

$$\mathsf{cap}(S^{\text{end}}_k) = \log_2 \delta(s).$$

Proof: First we note that by requiring $|s| \geqslant k$ we avoid the degenerate case of S^{end}_k containing only s. We further note that, by the definition of the replication functions,

$$R(x) = R(T^{\text{end}}_{i,k}(x))$$

for all non-negative integers i and k, and thus, all the strings in S^{end}_k have the same alpha-representation. Thus, trivially,

$$\mathsf{cap}(S^{\text{end}}_k) \leqslant \log_2 \delta(s).$$

We now turn to prove the inequality in the other direction. We contend that given a string $x \in \Sigma^*$, $|x| \geqslant k$, and some string $w \in \Sigma^k$, $R(w) \subseteq R(x)$, with at most $2k$ replication steps we can obtain from x a string $y \in \Sigma^*$ ending with w, i.e., $y = vw$.

As a first step, we replicate the prefix of x, i.e., if $x = uv$, $|u| = k$, then

$$x' = T^{\text{end}}_{0,k}(x) = uvu.$$

By doing so we ensure that for any symbol $a \in R(x)$ there is a k-substring of x' starting with a, and a k-substring of x' ending with a.

Let us now denote the symbols of w as $w = w_1w_2\dots w_k$, $w_i \in \Sigma$. Assume that the k-substring of x' starting at position i_1 ends with w_1. We form

$$x_1 = T^{\text{end}}_{i_1-1,k}(x')$$

whose 1-suffix is just w_1. Next, assume the k-substring of x' starting at position i_2 starts with w_2. Note that x' is a prefix of x_1. We form

$$x_2 = T^{\text{end}}_{|x_1|-k+1,k}\left(T^{\text{end}}_{i_2-1,k}(x_1)\right).$$

It easy to verify x_2 has a 2-suffix of w_1w_2. Continuing in the same way, let i_j be starting position of a k-substring of x' starting with w_j. We form

$$x_j = T^{\text{end}}_{|x_{j-1}|-k+1,k}\left(T^{\text{end}}_{i_j-1,k}(x_{j-1})\right),$$

for $j = 3, \dots, k$. Note that x_j has a j-suffix $w_1, \dots, w_j$.

It follows that after $2k$ replication steps we can obtain from any such x a string with any given k-suffix w, provided $R(w) \subseteq R(x)$. Thus, from the initial string s, we can obtain

a string s' with all of the strings of $R(s)^k$ appearing as k-substrings, using at most $2k\delta(s)^k$ replication steps[1], i.e.,

$$|s'| \leqslant |s| + 2k^2\delta(s)^k.$$

After having obtained s', each replication may replicate any of the k-strings in $R(s)^k$ in a single operation. Thus, for all $n = |s'| + tk$, t a non-negative integer, the number of distinct strings in S_k^{end} is bounded from below by

$$N_{S_k^{\text{end}}}(n) \geqslant \delta(s)^{n-|s'|}.$$

Since $|s'|$ is a constant, we have

$$\mathsf{cap}(S_k^{\text{end}}) \geqslant \log_2 \delta(s).$$

■

The following is an obvious corollary.

Theorem 2. *Let Σ be any finite alphabet, $k \geqslant 1$ any integer, and $s \in \Sigma^*$, $|s| \geqslant k$. Then for $S_{\geqslant k}^{\text{end}} = (\Sigma, s, \mathcal{T}_{\geqslant k}^{\text{end}})$,*

$$\mathsf{cap}(S_{\geqslant k}^{\text{end}}) = \mathsf{cap}(S_k^{\text{end}}) = \log_2 \delta(s).$$

Proof: Since for all $n \geqslant k$,

$$N_{S_k^{\text{end}}}(n) \leqslant N_{S_{\geqslant k}^{\text{end}}}(n) \leqslant \delta(s)^n,$$

the claim follows. ■

IV. Tandem Replication

We now consider different replication rules, $T_{i,k}^{\tan} : \Sigma^* \to \Sigma^*$, defined by

$$T_{i,k}^{\tan}(x) = \begin{cases} uvvw & \text{if } x = uvw,\ |u| = i,\ |v| = k \\ x & \text{otherwise.} \end{cases}$$

We also define the sets

$$\mathcal{T}_k^{\tan} = \left\{ T_{i,k}^{\tan} \mid i \geqslant 0 \right\}$$
$$\mathcal{T}_{\geqslant k}^{\tan} = \left\{ T_{i,k'}^{\tan} \mid i \geqslant 0, k' \geqslant k \right\}$$

Unlike the end replication discussed in the previous section, tandem replication takes a k-substring and replicates it adjacent to itself in the string. Also, the capacity of tandem-replication systems is in complete contrast to end-replication systems.

Theorem 3. *Let Σ be any finite alphabet, k any positive integer, and $s \in \Sigma^*$, with $|s| \geqslant k$. Then for $S_k^{\tan} = (\Sigma, s, \mathcal{T}_k^{\tan})$,*

$$\mathsf{cap}(S_k^{\tan}) = 0.$$

Proof: Consider any n-string $x \in \Sigma^*$, $|x| \geqslant k$. Instead of viewing $x = x_1x_2\ldots x_n$ as a sequence of n symbols from Σ, we can, by abuse of notation, view it as a sequence of $n - k + 1$ overlapping k-substrings $x = x'_1x'_2\ldots x'_{n-k+1}$, where

$$x'_i = x_ix_{i+1}\ldots x_{i+k-1}.$$

[1]This bound may be improved, but this will not affect the capacity calculation.

For a k-string $y = y_1y_2\ldots y_k$, $y_i \in \Sigma$, its cyclic shift by one position is denoted by

$$Ey = y_2y_3\ldots y_ky_1.$$

A cyclic shift by j positions is denoted by

$$E^jy = y_{j+1}y_{j+2}\ldots y_ky_1y_2\ldots y_j.$$

We say two k-strings, $y, z \in \Sigma^k$, are cyclically equivalent if

$$y = E^jz,$$

for some integer j. Clearly this is an equivalence relation. Let $\phi(y)$ denote the equivalence class of y. If y and z are cyclically equivalent, then $\phi(y) = \phi(z)$.

We now define

$$\Phi(x) = \phi(x'_1)\phi(x'_2)\ldots\phi(x'_{n-k+1}),$$

i.e., $\Phi(x)$ is the image of the overlapping k-substrings of x under ϕ. We also observe that knowing x'_1 and $\Phi(x)$ enables a full reconstruction of x.

At this point we turn to consider the effect of the replication $T_{i,k}^{\tan}$ on a string $x \in \Sigma^*$, $|x| \geqslant k$. When viewed as a sequence of overlapping k-substrings, as defined above,

$$T_{i,k}^{\tan}(x) = x'_1\ldots x'_{i-1}x'_iEx'_iE^2x'_i\ldots E^{k-1}x'_ix'_ix'_{i+1}\ldots x'_{n-k+1}.$$

Since $\phi(x'_i) = \phi(E^j(x'_i))$ for all j, we have

$$\begin{aligned}\Phi(T_{i,k}^{\tan}(x)) = {} & \phi(x'_1)\ldots\phi(x'_{i-1}) \\ & \phi(x'_i)\phi(x'_i)\ldots\phi(x'_i) \\ & \phi(x'_{i+1})\ldots\phi(x'_{n-k+1}),\end{aligned}$$

where $\phi(x'_i)$ appears $k+1$ consecutive times.

Thus, we may think of $\phi(x'_i)$ as a bin, and the action of $T_{i,k}^{\tan}$ as throwing k balls into the bin $\phi(x'_i)$. The number of bins does not change throughout the process, and is equal to one more than the number of times $\phi(x'_i) \neq \phi(x'_{i+1})$, where $x = s$ is the original string. If b is the number of bins defined by s, then the number of strings obtained by m replications is exactly $\binom{b+m-1}{b-1}$. Since this number grows only polynomially in the length of the resulting string, we have

$$\mathsf{cap}(S_k^{\tan}) = 0.$$

■

When considering $S_{\geqslant k}^{\tan} = (\Sigma, s, T_{\geqslant k}^{\tan})$ the situation appears to be harder to analyze.

Theorem 4. *For any finite alphabet Σ, and any string $s \in \Sigma^*$ of nontrivial alpha-diversity, $\delta(s) \geqslant 2$, we have*

$$\mathsf{cap}(S_{\geqslant 1}^{\tan}) \geqslant \log_2(r+1),$$

where r is the largest (real) root of the polynomial

$$f(x) = x^{\delta(s)} - \sum_{i=0}^{\delta(s)-2} x^i.$$

Proof: The proof strategy is the following: we shall show that $S_{\geqslant 1}^{\tan}$ contains, among other things, a regular language. The

capacity of that regular language will serve as the lower bound we claim.

For the first phase of the proof, assume $i_1 < i_2 < \cdots < i_{\delta(s)}$ are the indices of $\delta(s)$ distinct alphabet symbols in s. We produce a sequence of strings, $s_0 = s, s_1, \ldots, s_{\delta(s)-1}$, defined iteratively by

$$s_j = T^{\tan}_{i_{\delta(s)-j}-1, i_{\delta(s)}-i_{\delta(s)-j}+j}(s_{j-1}),$$

for $j = 1, 2, \ldots, \delta(s) - 1$. After this set of steps, the $\delta(s)$-substring starting at position $i_{\delta(s)}$ of $s_{\delta(s)-1}$ contains $\delta(s)$ distinct symbols. In what follows we will only use these symbols for replication, and thus, the constant amount of other symbols in $s_{\delta(s)-1}$ does not affect the capacity calculation. Thus, for ease of presentation we shall assume from now on that $|s| = \delta(s)$, i.e., the initial string contains no repeated symbol from the alphabet. Furthermore, without loss of generality, let us assume these symbols are $a_{\delta(s)}, a_{\delta(s)-1}, \ldots, a_1$, in this order.

We now perform the following iterations: In iteration i, where $i = \delta(s), \delta(s) - 1, \ldots, 2$, we replicate i-substrings equal to $a_i a_{i-1} \ldots a_2 a_1$. As a final iteration, we may replicate 1-substrings without constraining their content. It is easy to verify the resulting strings form the following regular language,

$$S = \left(a^+_{\delta(s)} \left(a^+_{\delta(s)-1} \left(\ldots \left(a^+_2 \left(a^+_1 \right)^+ \right)^+ \right)^+ \right)^+ \right)^+ .$$

The construction process implies $S \subseteq S^{\tan}_{\geqslant 1}$.

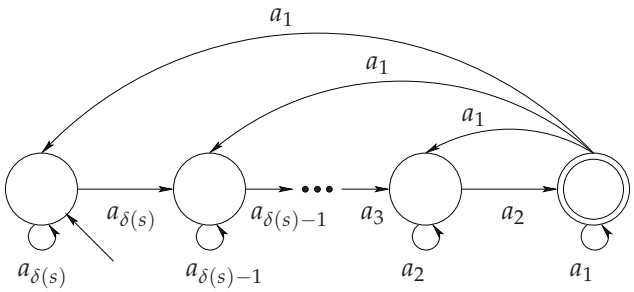

Figure 1. The finite-state automaton accepting the regular language used in the proof of Theorem 4.

The finite-state automaton accepting S is depicted in Figure 1. The graph is primitive and lossless, and thus, for the purpose of calculating the capacity, instead of counting the number of length n words in S, we can count the number of length n paths in the automaton graph $\mathcal{G}$ (see [4], [9]). By Perron-Frobenius theory,

$$\mathsf{cap}(S^{\tan}_{\geqslant 1}) \geqslant \mathsf{cap}(S) = \log_2 \lambda(A_{\mathcal{G}}),$$

where $\lambda(A_{\mathcal{G}})$ is the largest magnitude of an eigenvalue of $A_{\mathcal{G}}$, and where $A_{\mathcal{G}}$ denotes the adjacency matrix of $\mathcal{G}$. We note that $A_{\mathcal{G}}$ is the $\delta(s) \times \delta(s)$ matrix

$$A_{\mathcal{G}} = \begin{pmatrix} 1 & 1 & & & & \\ & 1 & 1 & & & \\ & & 1 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & 1 & 1 \\ 1 & 1 & 1 & \ldots & 1 & 1 \end{pmatrix},$$

and its largest eigenvalue is the largest real root of

$$\det(\lambda I - A_{\mathcal{G}}) = (\lambda - 1)^{\delta(s)} - \sum_{i=0}^{\delta(s)-2} (\lambda - 1)^i.$$

Setting $x = \lambda - 1$ we obtain the desired result. ∎

At least in one case, the bound of Theorem 4 is attained with equality, as is shown in the following corollary.

Corollary 5. *For $\Sigma = \{0, 1\}$, and $s \in \Sigma^*$ with $\delta(s) = 2$ we have*

$$\mathsf{cap}(S^{\tan}_{\geqslant 1}) = 1.$$

Proof: By applying Theorem 4 we get

$$\mathsf{cap}(S^{\tan}_{\geqslant 1}) \geqslant 1.$$

We also have the trivial upper bound

$$\mathsf{cap}(S^{\tan}_{\geqslant 1}) \leqslant \log_2 |\Sigma| = 1,$$

which completes the proof. ∎

For $S^{\tan}_{\geqslant k}$ and general k, we claim a weaker result, that is provided in the following theorem.

Theorem 6. *For any finite alphabet Σ, and any binary string $s \in \Sigma^*$, $|s| \geqslant k$, of nontrivial alpha-diversity, $\delta(s) \geqslant 2$, we have*

$$\mathsf{cap}(S^{\tan}_{\geqslant k}) \geqslant \log_2 r > 0,$$

where r is the largest root of the polynomial

$$f(x) = x^{k+1} - x - 1.$$

Proof: The proof strategy is, again, to find a regular language that is a subset of $S^{\tan}_{\geqslant k}$ and use its capacity as a lower bound. We start with the following preparation, by performing the following k replications,

$$s' = T^{\tan}_{0,2k-1} \left(\ldots \left(T^{\tan}_{0,k+1} \left(T^{\tan}_{0,k}(s) \right) \right) \right).$$

If we denote $s' = s'_1 s'_2 \ldots$, where $s'_i \in \Sigma$, then it is easy to verify that

$$s'_{k+1} = s'_{k+2} = \cdots = s'_{2k} = s'_1.$$

Since $\delta(s') = \delta(s) \geqslant 2$, this run of at least k consecutive equal symbols, must end. Without loss of generality, assume $0, 1 \in \Sigma$, and (possibly after an appropriate relabeling of the symbol names) either $0^k 1$ or 10^k form a substring of s'. We shall assume the former, and the proof for the latter case is similar. We ignore the rest of the symbols, as they will not affect the capacity. Thus, we may proceed as if the initial string s is $0^k 1$.

We now generate more strings by replicating only substrings of the form 0^k1 or $0^{k-1}1$. The resulting set of strings contains the regular language

$$S = \left(\left(0^k1\right)^+ \left(0^{k-1}1\right)^+ \right)^+.$$

We can follow the same steps as in the proof of Theorem 4 in order to find the capacity of S. It is given by the base-2 logarithm of the largest real solution for the equation

$$x^{-(k+1)} + x^{-k} = 1.$$

By rearranging, we get the claim. It is also easy to verify that the claimed r satisfies $r > 1$, and so the capacity is strictly positive. ∎

V. Reversed Tandem Replication

Consider the reversed tandem replication rule $T^{\mathrm{rt}}_{i,k} : \Sigma^* \to \Sigma^*$ defined as

$$T^{\mathrm{rt}}_{i,k}(x) = \begin{cases} uvv^Rw & \text{if } x = uvw,\ |u| = i,\ |v| = k, \\ x & \text{otherwise,} \end{cases}$$

where y^R is the reverse of y, i.e., $y^R = y_m y_{m-1} \dots y_1$ for a sequence $y = y_1 y_2 \dots y_m \in \Sigma^*$. Furthermore, let

$$\mathcal{T}^{\mathrm{rt}}_k = \left\{ T^{\mathrm{rt}}_{i,k} \;\middle|\; i \geqslant 0 \right\}.$$

and use $S^{\mathrm{rt}}_k = (\Sigma, s, \mathcal{T}^{\mathrm{rt}}_k)$. Since the starting string s will play a crucial role, we shall often use the notation $S^{\mathrm{rt}}_k(s)$.

Lemma 7. *Let $s \in \Sigma^k$ such that $s \neq s^R$. Then*

$$\mathsf{cap}(S^{\mathrm{rt}}_k(s)) \geqslant \frac{1}{k}.$$

Proof: By repeatedly applying replication to the last block of k symbols, we can create any sequence of alternating blocks s and s^R, starting with s. To extend any run of s (resp. s^R), except the first one, we can apply replication to the last block of the previous run, which is an s^R block (resp. s). Thus, the regular language

$$S = ss^R \left\{s, s^R\right\}^*,$$

satisfies $S \subseteq S^{\mathrm{rt}}_k(s)$. Since $s \neq s^R$, we easily see that

$$\mathsf{cap}(S^{\mathrm{rt}}_k(s)) \geqslant \mathsf{cap}(S) = \frac{1}{k}.$$

∎

Note that the requirement that $s \neq s^R$ implies that $k \geqslant 2$.

The following theorem states that the capacity of reversed tandem replication is positive except in trivial cases.

Theorem 8. *For any $s \in \Sigma^*$, $|s| \geqslant k$, we have $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) = 0$ if and only if $\delta(s) = 1$.*

Proof: It is clear that if $\delta(s) = 1$, then $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) = 0$. For the other direction, suppose that $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) = 0$. We show that $\delta(s) = 1$. We first prove this for $|s| = k$.

Denote $s = s_1 s_2 \dots s_k$, with $s_i \in \Sigma$. Since $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) = 0$, by Lemma 7, we have that $s = s^R$, or equivalently,

$$s_i = s_{k+1-i}, \quad \forall i \in [k]. \tag{1}$$

From $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) = 0$, it also follows that $\mathsf{cap}(S^{\mathrm{rt}}_k(ss^R)) = 0$, which in turn implies that $\mathsf{cap}(S^{\mathrm{rt}}_k(s_2 s_3 \dots s_k s_k)) = 0$. Hence,

$$\begin{aligned} s_2 &= s_k, \\ s_{i+1} &= s_{k+2-i}, \quad \forall 2 \leqslant i \leqslant k-1. \end{aligned}$$

or equivalently,

$$s_2 = s_k, \tag{2}$$
$$s_{i+2} = s_{k+1-i}, \quad \forall i \in [k-2]. \tag{3}$$

From (1) and (3), it follows that

$$s_i = s_{i+2}, \quad \forall i \in [k-2]. \tag{4}$$

It is also true that $s_1 = s_2$ since $s_1 = s_k$ from (1) and $s_2 = s_k$ from (2). The expressions (4) and $s_1 = s_2$ prove that $\delta(s) = 1$.

Finally, let $s \in \Sigma^*$ be such that $|s| \geqslant k$. If s' is a k-substring of s, then obviously

$$\mathsf{cap}(S^{\mathrm{rt}}_k(s)) \geqslant \mathsf{cap}(S^{\mathrm{rt}}_k(s')).$$

Since we have $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) = 0$, then $\mathsf{cap}(S^{\mathrm{rt}}_k(s')) = 0$, and using the above proof for length k strings, we get $\delta(s') = 1$. Since this is true for every k-substring s' of s, we must have $\delta(s) = 1$. ∎

In Theorem 10, we show that in determining the capacity of a system $S^{\mathrm{rt}}_k(s)$, only $\delta(s)$ is important and not the actual sequence s. The idea behind the proof is that any other finite sequence with alphabet $R(s)$ appears as a substring of some sequence in $S^{\mathrm{rt}}_k(s)$. To show this, we use the following lemma in the proof of Theorem 10.

Lemma 9. *For any $x, y \in \Sigma^*$, with $|y| \geqslant k$, if for all $a \in \Sigma$, $n_y(a) \geqslant n_x(a)$, then x is a suffix of some sequence in $S^{\mathrm{rt}}_k(y)$.*

Proof: Since we can increase the length of y by applying the function $T^{\mathrm{rt}}_{0,k}$, while maintaining $n_y(a) \geqslant n_x(a)$ for all $a \in \Sigma$, we assume without loss of generality that $|y| \geqslant 2k$. We also assume $|x| > 0$, or else the claim is trivial.

Suppose that the *last* symbol of x is a. We construct a sequence y'' from y using the functions $\mathcal{T}^{\mathrm{rt}}_k$ such that a is the last symbol of y'', i.e., a is "pushed" to the end. Let i be such that $y_i = a$. Consider the conditions

$$i \geqslant k, \qquad |y| - i \geqslant k.$$

At most one of the two conditions does not hold. If the former does not hold, let $y' = T^{\mathrm{rt}}_{0,k}(y)$. There is a copy of a at position $i' = 2k - i + 1$ in y', i.e., $y'_{i'} = a$. We have $i' \geqslant k$ and $|y'| - i' \geqslant 3k - (2k - i + 1) \geqslant k$. If the latter does not hold, let $y' = T^{\mathrm{rt}}_{i-k,k}(y)$ and $i' = i$. If both conditions hold, let $y' = y$ and $i' = i$. We thus have $y'_{i'} = a$ with $i' \geqslant k$ and $|y'| - i' \geqslant k$. The significance of these conditions is that they enable us to replicate blocks of length k containing a without the need to concern ourselves with the boundaries of the sequence.

Let $|y'| - i' = q(k-1) + r$ such that q and r are integers with $q \geqslant 1$ and $0 \leqslant r < k - 1$.

First, suppose k is even. We let $y'' = T^{\mathrm{rt}}_{i'-k/2,k}(y')$. Now there is a copy of a in y'' at position $i'' = i' + k + 1$. The distance of this copy from the end of y'' is

$$|y''| - i'' = |y'| + k - (i' + k + 1) = |y'| - i' - 1.$$

Hence, the distance is decreased by one, compared with y'. We repeat the same procedure and update y'' and i'' as

$$y'' \leftarrow T^{\mathrm{rt}}_{i''-k/2,k}(y''),$$
$$i'' \leftarrow i'' + k + 1,$$

until we have $|y''| - i'' = q(k-1)$. At this point we switch to repeating

$$y'' \leftarrow T^{\mathrm{rt}}_{i''-1,k}(y''), \tag{5}$$
$$i'' \leftarrow i'' + 2k - 1, \tag{6}$$

until a becomes the last symbol of y''.

Next, suppose that k is odd and r is even. We let $y'' = T^{\mathrm{rt}}_{i'-(k-1)/2}(y')$. Now there is a copy of a in y'' at position $i'' = i' + k + 2$. The distance of this copy from the end of y'' is

$$|y''| - i'' = |y'| + k - (i' + k + 2) = |y'| - i' - 2.$$

The distance is thus decreased by two, compared with y'. Since r is even, by repeating the same procedure and updating y'' and i'', we can have $|y''| - i'' = q(k-1)$. We then repeat (5) and (6) until a becomes the last symbol of y''.

Finally, suppose that k and r are both odd. Let $y'' = T^{\mathrm{rt}}_{i-1,k}(y')$. There is a copy of a in y'' at position $i'' = i'$. The distance of this copy from the end of y'' is $|y''| - i'' = |y'| + k - i$. Let $|y''| - i'' = q'(k-1) + r'$ where q' and r' are integers with $q' \geqslant 1$ and $0 \leqslant r' < k-1$. We thus have

$$r' = r + 1 + (q + 1 - q')(k-1).$$

Since $k-1$ is even and r is odd, we find that r' is even. We can then proceed as in the previous case in which k is odd and r is even.

We have shown that any symbol present in y can be "pushed" to the end position. We repeatedly apply the same argument by disregarding the last element of y'' and pushing the next appropriate element to the end position. The final result is a sequence in $S^{\mathrm{rt}}_k(y)$ which ends with x. ∎

Theorem 10. *For all $s \in \Sigma^*$, $|s| \geqslant k$, $\mathsf{cap}(S^{\mathrm{rt}}_k(s))$ depends on s only through $\delta(s)$.*

Proof: Consider two sequences $s, t \in \Sigma^*$, $|s|, |t| \geqslant k$, such that $\delta(s) = \delta(t)$. Since the identity of the symbols is irrelevant to the capacity, we may assume that $R(s) = R(t)$. By appropriate replications, it is easy to find a sequence $t' \in S^{\mathrm{rt}}_k(t)$ such that for all $a \in \Sigma$, we have $n_{t'}(a) \geqslant n_s(a)$. We then apply Lemma 9 and show that s is a substring of some sequence $t'' \in S^{\mathrm{rt}}_k(t)$. Hence,

$$\mathsf{cap}(S^{\mathrm{rt}}_k(s)) \leqslant \mathsf{cap}(S^{\mathrm{rt}}_k(t'')) \leqslant \mathsf{cap}(S^{\mathrm{rt}}_k(t)).$$

Similarly, we can show that $\mathsf{cap}(S^{\mathrm{rt}}_k(t)) \leqslant \mathsf{cap}(S^{\mathrm{rt}}_k(s))$. Hence, $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) = \mathsf{cap}(S^{\mathrm{rt}}_k(t))$. ∎

TABLE I
NUMERICAL RESULTS FOR REVERSED TANDEM REPLICATION

$s = 01, k = 2$	n	2	4	6	8	10	12	14
	$N(n)$	1	1	3	10	37	145	584
$s = 010, k = 3$	n	3	6	9	12	15	18	21
	$N(n)$	1	1	3	14	78	467	2894
$s = 012, k = 3$	n	3	6	9	12	15	18	21
	$N(n)$	1	1	4	25	182	1423	11577

Example 11. *Suppose s is a string of length k such that $s = s^R$. We show that for positive integers p and q, we have*

$$N_{S^{\mathrm{rt}}_k(s)}(pqk) \geqslant \left(N_{S^{\mathrm{rt}}_k(s)}(pk)\right)^q. \tag{7}$$

To see this, note that to generate sequences of length pqk, we can first generate a sequence of length qk consisting of q copies of s, and then from each of these copies, generate a sequence of length pk. It is clear that (7) also holds for the case in which s is equal to a relabeling of s^R, where the relabeling map is bijective, e.g., $s = 012$. If we let $q \to \infty$ in (7), we find that

$$\mathsf{cap}\left(S^{\mathrm{rt}}_k(s)\right) \geqslant \frac{\log_2 N_{S^{\mathrm{rt}}_k(s)}(pk)}{pk}. \tag{8}$$

Using a computer, we obtain Table I for the given values of s and k, and then use (8) to find the following lower bounds on the capacity,

$$\mathsf{cap}\left(S^{\mathrm{rt}}_2(01)\right) \geqslant \frac{\log_2 584}{14} \geqslant 0.65,$$
$$\mathsf{cap}\left(S^{\mathrm{rt}}_3(010)\right) \geqslant \frac{\log_2 2894}{21} \geqslant 0.54,$$
$$\mathsf{cap}\left(S^{\mathrm{rt}}_3(012)\right) \geqslant \frac{\log_2 11577}{21} \geqslant 0.64.$$

□

VI. REPLICATION WITH A GAP

Consider the replication-with-a-gap rule $T^{\mathrm{gap}}_{i,k,k'} : \Sigma^* \to \Sigma^*$ defined as

$$T^{\mathrm{gap}}_{i,k,k'}(x) = \begin{cases} uvwvz, & \text{if } x = uvwz,\ |u| = i, \\ & |v| = k,\ |w| = k', \\ x, & \text{otherwise.} \end{cases}$$

Furthermore, we let

$$\mathcal{T}^{\mathrm{gap}}_{k,k'} = \left\{ T^{\mathrm{gap}}_{i,k,k'} \;\middle|\; i \geqslant 0 \right\},$$

and use $S^{\mathrm{gap}}_{k,k'} = (\Sigma, s, \mathcal{T}^{\mathrm{gap}}_{k,k'})$, for some $s \in \Sigma^*$. We may also use $S^{\mathrm{gap}}_{k,k'}(s)$ to denote the aforementioned string system. To avoid trivialities, throughout this section, we assume $k, k' \geqslant 1$.

For a sequence $s = s_1 s_2 \ldots$, with $s_i \in \Sigma$, we conveniently denote the substring starting at position i and of length k as $s_{i,k} = s_i s_{i+1} \ldots s_{i+k-1}$. Furthermore, for two sequences of equal length, $s, s' \in \Sigma^k$, we denote their Hamming distance as $d_H(s, s')$, which is the number of coordinates in which s and s' disagree.

Lemma 12. *For all $s \in \Sigma^*$ such that $|s| \geqslant k + k'$, we have*

$$\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(s)) \geqslant \frac{1}{k}\log_2\left(1 + d_H\left(s_{1,k}, (s^2)_{k+1,k}\right)\right).$$

Proof: The proof considers two cases: either $k \geqslant k'$, or $k < k'$. We prove the former. The proof for the latter is similar. It also suffices to consider only $|s| = k + k'$, since for longer strings we can simply ignore the extra symbols.

For simplicity of notation, let $s = x_1 \dots x_k y_1 \dots y_{k'}$, where $x_i, y_i \in \Sigma$. We initially apply $T^{\mathrm{gap}}_{0,k,k'}$ to s and obtain

$$s' = T^{\mathrm{gap}}_{0,k,k'}(s) = x_1 \dots x_k y_1 \dots y_{k'} x_1 \dots x_k.$$

We then apply $T^{\mathrm{gap}}_{i,k,k'}$ to s', for all $0 \leqslant i \leqslant k$, and get the following list of results:

$$\begin{array}{l}
x_1 \dots x_k\ y_1 \dots y_{k'}\ x_1 \dots x_k\ x_1 x_2 \dots x_{k'} x_{k'+1} \dots x_k \\
x_1 \dots x_k\ y_1 \dots y_{k'}\ x_1 \dots x_k\ y_1 x_2 \dots x_{k'} x_{k'+1} \dots x_k \\
\vdots \\
x_1 \dots x_k\ y_1 \dots y_{k'}\ x_1 \dots x_k\ y_1 y_2 \dots y_{k'} x_{k'+1} \dots x_k \\
x_1 \dots x_k\ y_1 \dots y_{k'}\ x_1 \dots x_k\ y_1 y_2 \dots y_{k'}\ x_1\ \dots\ x_k \\
\vdots \\
x_1 \dots x_k\ y_1 \dots y_{k'}\ x_1 \dots x_k\ y_1 y_2 \dots y_{k'} x_1 \dots x_{k-k'}
\end{array}$$

where the five explicitly stated sequences correspond to $i = 0, 1, k', k'+1, k$. From these results, it is clear that we have $1 + d_H(s_{1,k}, (s^2)_{k+1,k})$ distinct sequences. Since the same operation can be repeated, i.e., apply $T^{\mathrm{gap}}_{i,k,k'}$ to s', for all $0 \leqslant i \leqslant k$, to all the distinct results of the previous round, the number of sequences in $S^{\mathrm{gap}}_{k,k'}$ with length $2k + k' + ik$ is at least

$$N_{S^{\mathrm{gap}}_{k,k'}}(2k + k' + ik) \geqslant \left(1 + d_H(s_{1,k}, (s^2)_{k+1,k})\right)^i.$$

This completes the proof. ∎

With an example, we show that the lower bound of Lemma 12 is sharp. Choose s as

$$s = a_1 \dots a_k b a_2 \dots a_k,$$

where $b \neq a_1$. Suppose $t \in S^{\mathrm{gap}}_{k,k}(s)$. Each k-substring $t_{(i-1)k+1,k}$, for nonnegative integers $i \leqslant |t|/k$, either equals $a_1 \dots a_k$ or $b a_2 \dots a_k$. Thus for a nonnegative integer j there are no more than 2^j sequence of length jk in $S^{\mathrm{gap}}_{k,k}(s)$. Hence,

$$\mathsf{cap}(S^{\mathrm{gap}}_{k,k}(s)) \leqslant \lim_{j\to\infty} \frac{\log_2 2^j}{jk} = \frac{1}{k}$$

which matches the lower bound given in Lemma 12, and so $\mathsf{cap}(S^{\mathrm{gap}}_{k,k}(s)) = \frac{1}{k}$.

The next corollary is an immediate result of the previous lemma.

Corollary 13. *Assume $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(s)) = 0$, where $s \in \Sigma^*$ and $|s| \geqslant k + k'$. For any $(k + k')$-substring of s, denoted $x_1 \dots x_k y_1 \dots y_{k'}$, with $x_i, y_i \in \Sigma$, we have*

$$\begin{array}{ll}
x_1 \dots x_k = y_1 \dots y_{k'} x_1 \dots x_{k-k'}, & \text{if } k > k', \\
x_1 \dots x_k = y_1 \dots y_k, & \text{if } k \leqslant k'.
\end{array}$$

This corollary is used in the following theorem.

Theorem 14. *For $s \in \Sigma^*$, $|s| \geqslant k + k'$, we have $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(s)) = 0$ if and only if s is periodic with period $\gcd(k, k')$.*

Proof: We start with the easy direction. Assume s is periodic with period $\gcd(k, k')$. Note that in this case $S^{\mathrm{gap}}_{k,k'}(s)$ contains only one sequence of length $ik + k'$ for each $i \geqslant 1$, which is itself a periodic extension of s. No other sequences appear in $S^{\mathrm{gap}}_{k,k'}(s)$. Thus, the capacity is 0.

We now turn to the other direction. Assume the capacity is 0. We further assume $s = x_1 \dots x_k y_1 \dots y_{k'}$, with $x_i, y_i \in \Sigma$, has length $k + k'$. The general case then follows easily. The proof in this direction is divided into two cases.

For the first case, let $k > k'$, and denote $k'' = k - k'$. We show that s is periodic with period $\gcd(k, k')$. From Corollary 13, it follows that $y_1 \dots y_{k'} = x_1 \dots x_{k'}$ so we can write $s = x_1 \dots x_k x_1 \dots x_{k'}$. Furthermore, said corollary implies that $x_i = x_{k'+i}$ for $i \in [k - k']$ and so $s = x_1 \dots x_{k'} x_1 \dots x_{k''} x_1 \dots x_{k'}$. By once applying the rule of $T^{\mathrm{gap}}_{0,k,k'}$ we obtain

$$t = x_1 \dots x_{k'}\ x_1 \dots x_{k''}\ x_1 \dots x_{k'}\ x_1 \dots x_{k'}\ x_1 \dots x_{k''}.$$

Now let us apply Corollary 13 to the substring $t' = x_1 \dots x_{k''}\ x_1 \dots x_{k'}\ x_1 \dots x_{k'}$ of t. Since $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(s)) = 0$, we must have $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(t)) = 0$, and obviously, also $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(t')) = 0$. Applying Corollary 13 to the last case of t', we get that

$$x_1 \dots x_{k''} x_1 \dots x_{k'} = x_1 \dots x_{k'} x_1 \dots x_{k''},$$

that is, the sequence $x_1 \dots x_{k''} x_1 \dots x_{k'}$, which has length k, equals itself when cyclically shifted by k'. Hence, it is periodic with period $\gcd(k, k')$ and thus s is periodic with the same period.

For the second case, let $k \leqslant k'$. Denote $x = x_1 \dots x_k$ and $y = y_1 \dots y_{k'}$, so $s = xy$. Find integers q and r such that $k' = qk + r$ and $0 \leqslant r < k$ and let t be the sequence obtained from s by $q + 1$ times applying $T^{\mathrm{gap}}_{0,k,k'}$, that is,

$$\begin{aligned}
t &= xyx^{q+1} \\
&= x_{1,k}\, y_{1,k}\, y_{k+1,k} \cdots y_{(q-1)k+1,k}\, y_{qk+1,r}\, (x_{1,k})^{q+1}.
\end{aligned}$$

Note that since $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(t)) = 0$, we also have $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(t')) = 0$ for any $(k + k')$-substring t' of t. Hence, we can apply Corollary 13 to any $(k + k')$-substring t' of t.

For $i = 0, 1, \dots, q - 1$, in that order, applying Corollary 13 to the $(k + k')$-substring $t_{ik+1,k+k'}$ implies that

$$x_{1,k} = y_{ik+1,k}. \tag{9}$$

Next, note that from (9), for the $(k + k')$-substring $t_{qk+1,k+k'}$, we have

$$\begin{aligned}
t_{qk+1,k+k'} &= y_{(q-1)k+1,k}\, y_{qk+1,r}\, (x_{1,k})^q \\
&= x_{1,k}\, y_{qk+1,r}\, (x_{1,k})^q.
\end{aligned}$$

By applying Corollary 13 to this sequence, we find

$$t_{qk+1,k+k'} = x_{1,k} x_{1,r} \left(x_{1,k}\right)^q .$$

Thus, we have

$$t = (x_{1,k})^{q+1}(x_{1,r})(x_{1,k})^{q+1}.$$

Finally, we apply Corollary 13 to the $(k+k')$-substring

$$t_{qk+r+1,k+k'} = x_{r+1} \cdots x_k \, x_1 \cdots x_r \, x_1 \cdots x_k$$

which shows that

$$x_{r+1} \cdots x_k \, x_1 \cdots x_r = x_1 \cdots x_k.$$

Since $x_1 \cdots x_k$ equals itself when cyclically shifted by r, it is periodic with period $\gcd(k,r) = \gcd(k,k')$. Hence t is periodic with the same period and so is s.

We have shown that for the special case of $|s| = k+k'$, if the capacity is zero, then s is periodic with period $\gcd(k,k')$. Now suppose $|s| > k+k'$ and that $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(s)) = 0$. Let $d = \gcd(k,k')$ and, for the moment, also suppose that d divides $|s|$. Let

$$C = \left\{ s_{id+1,k+k'} : 0 \leqslant i \leqslant \frac{|s| - (k+k')}{d} \right\}$$

be a set of $(k+k')$-substrings of s that cover s and each consecutive pair overlap in d positions. Since the capacity for each of these $(k+k')$-substrings is also zero, they are periodic with period d. Because of their overlaps and the fact that they cover s, it follows that s is also periodic with period d. To complete the proof it remains to consider the case in which d does not divide $|s|$. In this case, we can repeat the same argument but with adding the substring $s_{|s|-(k+k')+1,(k+k')}$ to the set C to ensure that s is covered by overlapping $(k+k')$-substrings. ∎

We now turn to discuss the dependence of $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(s))$ on s. For a sequence $x \in \Sigma^*$ and two symbols $a, b \in R(x)$, let

$$\Delta_x(a,b) = \left\{ j \;\middle|\; \exists i, x_i = a, x_{i+j} = b \right\},$$

be the set of the differences of positions of a and b in x. Furthermore, let

$$\rho_{x,\ell}(a,b) = \left\{ (j \bmod \ell) \mid j \in \Delta_x(a,b) \right\}.$$

Lemma 15. *Let Σ be some finite alphabet, $d > 0$ an integer, and $D \subset \{0,1,\ldots,d-1\}$ some subset, $|D| < d$. Consider the constrained system $S \subseteq \Sigma^*$ such that for every $x \in S$, and every two symbols $a, b \in \Sigma$ (not necessarily distinct), $\rho_{x,d}(a,b) \subseteq D$. Then $\mathsf{cap}(S) < \log_2 |\Sigma|$.*

Proof: We begin by constructing a De-Bruijn graph of order $d+1$ over Σ, $\mathcal{G}''(V'', E'')$, defined in the following way. We set $V'' = \Sigma^{d+1}$, and a directed edge connects $v = v_1 \ldots v_{d+1} \in V''$ and $v' = v'_1 \ldots v'_{d+1} \in V''$, if $v'_i = v_{i+1}$ for all $1 \leqslant i \leqslant d$. That edge has label $v'_{d+1} \in \Sigma$. The graph is regular with out-degree $|\Sigma|$. Clearly the set of finite strings read along paths taken in $\mathcal{G}''$ is simply $S'' = \Sigma^*$. In particular, by Perron-Frobenius theory, if $A_{\mathcal{G}''}$ is the adjacency matrix of $\mathcal{G}''$, since $\mathcal{G}''$ is clearly primitive,

$$\mathsf{cap}(S'') = \log_2 \lambda(A_{\mathcal{G}''}) = \log_2 |\Sigma| .$$

As the next step, we construct a graph $\mathcal{G}'(V', E')$ from $\mathcal{G}''(V'', E'')$ by setting $V' = V''$, and removing all edges $v \to u$, such that

$$\rho_{v,d}(a,b) \cup \rho_{u,d}(a,b) \not\subseteq D,$$

for some $a, b \in \Sigma$. The labels of the surviving edges remain the same. We define S' to be the set of strings read from finite paths in $\mathcal{G}'$. Since $|D| < d$, $A_{\mathcal{G}'}$ is obtained from $A_{\mathcal{G}''}$ be changing at least one entry from 1 to 0. By Perron-Frobenius theory,

$$\mathsf{cap}(S') \leqslant \log_2 \lambda(A_{\mathcal{G}'}) < \log_2 \lambda(A_{\mathcal{G}''}) = \log_2 |\Sigma| .$$

Finally, since it is clear that $S \subseteq S'$, we get

$$\mathsf{cap}(S) \leqslant \mathsf{cap}(S') < \log_2 |\Sigma| ,$$

as claimed. ∎

Using Lemma 15 we obtain the following theorem.

Theorem 16. *Let $s \in \Sigma^*$ have length at least $k+k'$ and denote $d = \gcd(k,k')$. If, for some $a, b \in R(s)$, we have $\left|\rho_{s,d}(a,b)\right| < d$ then $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(s)) < \log_2 \delta(s)$.*

Proof: We observe that for any $x, x' \in S^{\mathrm{gap}}_{k,k'}(s)$, and for $a, b \in R(s)$, we have

$$\rho_{x,d}(a,b) = \rho_{x',d}(a,b) ,$$

where $d = \gcd(k,k')$. This can be easily seen by noting that any function in $\mathcal{T}^{\mathrm{gap}}_{k,k'}$ changes the differences between positions of two elements by a linear combination of k and k'. We then apply Lemma 15. ∎

Theorem 17. *For $s \in \Sigma^*$ with $|s| \geqslant k+k'$, if $\gcd(k,k') = 1$, then $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(s))$ depends on s only through $\delta(s)$.*

Proof: The proof is similar to that of Theorem 10. In that light, it suffices to show that in a sequence $y \in \Sigma^*$ of length $m \geqslant k+k'$, a symbol $a \in R(y)$ can be "pushed" to the end. That is, we can find a sequence $y'' \in S^{\mathrm{gap}}_{k,k'}(y)$ that ends with a.

Suppose a is in position i in y. Without loss of generality (similar to Lemma 9), we may assume $i > k$ and $m - i \geqslant k' - 1$.

Let $y' = T^{\mathrm{gap}}_{i-1,k,k'}(y)$. There is a copy of a at position $i' = i$ whose distance from the end of y' is $|y'| - i' = k + m - i$ and this is an increase of size k compared to y. We update y' as $y' \leftarrow T^{\mathrm{gap}}_{i'-1,k,k'}(y')$. In each step, the distance of a at position i' from the end of y' increases by k. We continue until we have $k' \mid |y'| - i'$. This eventually happens as $\gcd(k,k') = 1$.

Now we let $y'' = T^{\mathrm{gap}}_{i'-k,k,k'}(y')$. There is a copy of a in y'' at position $i'' = i' + k + k'$. The distance of this copy of a from the end of y'' is $|y''| - i'' = |y'| - i' - k'$. Thus the distance is decreased by k'. We update y'' and i'' as $y'' \leftarrow T^{\mathrm{gap}}_{i''-k,k,k'}(y'')$ and $i'' \leftarrow i'' + k + k'$. We continue until a is the last element of y''. The rest of the argument follows along the same lines as those of Lemma 9 and Theorem 10. ∎

REFERENCES

[1] J. Dassow, V. Mitrana, and G. Paun, "On the regularity of duplication closure," *Bulletin of the EATCS*, vol. 69, pp. 133–136, 1999.

[2] J. Dassow, V. Mitrana, and A. Salomaa, "Operations and language generating devices suggested by the genome evolution," *Theoretical Computer Science*, vol. 270, no. 1, pp. 701–738, 2002.

[3] J. W. Fondon and H. R. Garner, "Molecular origins of rapid and continuous morphological evolution," *Proceedings of the National Academy of Sciences*, vol. 101, no. 52, pp. 18 058–18 063, 2004.

[4] K. A. S. Immink, *Codes for Mass Data Storage Systems*. Shannon Foundation Publishers, 2004.

[5] 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.

[6] P. Leupold, C. Martín-Vide, and V. Mitrana, "Uniformly bounded duplication languages," *Discrete Applied Mathematics*, vol. 146, no. 3, pp. 301–310, 2005.

[7] P. Leupold, V. Mitrana, and J. M. Sempere, "Formal languages arising from gene repeated duplication," in *Aspects of Molecular Computing*. Springer, 2004, pp. 297–308.

[8] G. Levinson and G. A. Gutman, "Slipped-strand mispairing: a major mechanism for DNA sequence evolution." *Molecular Biology and Evolution*, vol. 4, no. 3, pp. 203–221, 1987.

[9] D. Lind and B. H. Marcus, *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, 1985.

[10] N. Mundy and A. J. Helbig, "Origin and evolution of tandem repeats in the mitochondrial DNA control region of shrikes (lanius spp.)," *Journal of Molecular Evolution*, vol. 59, no. 2, pp. 250–257, 2004.

[11] K. Usdin, "The biological effects of simple tandem repeats: lessons from the repeat expansion diseases," *Genome research*, vol. 18, no. 7, pp. 1011–1019, 2008.