The Capacity of String-Duplication 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 duplication rules, including those resembling genomic duplication processes. In other words, our goal is to find out the capacity, or the expressive power, of these string-duplication systems. Our results include the exact capacities, and bounds on the capacities, of four fundamental string-duplication systems.**

I. Introduction

More than 50% of the human genome consists of *repeated sequences* [6]. 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 [6].

A second type of repeats are *tandem repeats*, generally thought to be caused by *slipped-strand mispairings* [11]. 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 [9]. 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 [12], and rapid morphological variation [4].

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 [6], [12].

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-duplication systems* with simple duplication rules including rules that resemble the repeat-producing genomic processes, namely duplication of transposons and duplication caused by slipped-strand mispairings. A string-duplication system, to be defined formally later, consists of a set of rewriting rules, an initial sequence, and all sequences that can be obtained by 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-duplication system, and thus illustrates the expressive power and the diversity of that system.

In this paper, we consider four duplication rules. The first is the *end duplication* 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 duplication* and allows a substring of length k to be duplicated 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 duplication*, which is similar to tandem duplication except that the copy is reversed before insertion. For instance, 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 *duplication 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 duplication has been already studied in a series of papers [1], [2], [7], [8]. 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).

978-1-4799-5186-4/14/$31.00 ©2014 IEEE

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 duplication rules. Due to lack of space, we omit some proofs or only provide their outline. Complete proofs can be found in the full version of this paper, available on arXiv [3].

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 $xy \in \Sigma^{n+m}$. For a positive integer m and a string s, s^m denotes the concatenation of m copies of s. 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 $n_s(a)$.

For $S \subseteq \Sigma^*$, we let $S^* = \{w_1w_2\dots w_m \mid w_i \in S, m \geqslant 0\}$, 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 $v \in \Sigma^*$.

A *string system* S is a subset $S \subseteq \Sigma^*$. For any integer n, we denote by $N_S(n)$ the number 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-duplication 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 duplication process), and $\mathcal{T}$ is a set of functions such that each $T \in \mathcal{T}$ is a mapping from Σ^* to Σ^* that defines a string-duplication rule. The resulting string system S, induced by $(\Sigma, s, \mathcal{T})$, is defined as the closure of the string-duplication 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 Duplication

We define the end-duplication 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:

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

Intuitively, in the end-duplication system, the transformations duplicate a substring of length k and append the duplicate 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 $S^{\text{end}}_k = (\Sigma, s, \mathcal{T}^{\text{end}}_k)$, we have $\mathsf{cap}(S^{\text{end}}_k) = \log_2 \delta(s)$.*

Proof: (Sketch). 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 duplication 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. It can be shown that after $2k$ duplication steps we can obtain from any $x \in \Sigma^*$, $|x| \geqslant k$, a string x' 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 $2k\delta(s)^k$ duplication steps[1], i.e.,

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

After having obtained s', each duplication may duplicate 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^{end}_k is bounded from below by

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

Since $|s'|$ is a constant, we have $\mathsf{cap}(S^{\text{end}}_k) \geqslant \log_2 \delta(s)$. ∎

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 $S^{\text{end}}_{\geqslant k} = (\Sigma, s, \mathcal{T}^{\text{end}}_{\geqslant k})$, we have $\mathsf{cap}(S^{\text{end}}_{\geqslant k}) = \mathsf{cap}(S^{\text{end}}_k) = \log_2 \delta(s)$.*

Proof: Since for all $n \geqslant k$, we have $N_{S^{\text{end}}_k}(n) \leqslant N_{S^{\text{end}}_{\geqslant k}}(n) \leqslant \delta(s)^n$, the claim follows. ∎

IV. Tandem Duplication

We now consider different duplication rules, $T^{\text{tan}}_{i,k} : \Sigma^* \to \Sigma^*$, defined by

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

We also define the sets

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

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

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

Theorem 3. *Let Σ be any finite alphabet, k any positive integer, and $s \in \Sigma^*$, with $|s| \geqslant k$. Then for $S_k^{\tan} = (\Sigma, s, \mathcal{T}_k^{\tan})$, we have* $\mathsf{cap}(S_k^{\tan}) = 0$.

Proof: Consider any n-string $x \in \Sigma^*$, $n \geqslant k$. Instead of viewing $x = x_1 x_2 \dots x_n$ 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 $x = x'_1 x'_2 \dots x'_{n-k+1}$, where

$$x'_i = x_i x_{i+1} \dots x_{i+k-1}.$$

For a k-string $y = y_1 y_2 \dots y_k$, $y_i \in \Sigma$, its cyclic shift by one position is denoted by $Ey = y_2 y_3 \dots y_k y_1$. A cyclic shift by j positions is denoted by

$$E^j y = y_{j+1} y_{j+2} \dots y_k y_1 y_2 \dots y_j.$$

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

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

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 duplication $T_{i,k}^{\tan}$ on a string $x \in \Sigma^*$, $|x| \geqslant k$. When viewed as a sequence of overlapping k-substrings, as defined above,

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

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

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

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 duplications 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 $i_1 < i_2 < \dots < i_{\delta(s)}$ are the indices of $\delta(s)$ distinct alphabet symbols in s. We produce a sequence of strings, $s_0, s_1, \dots, s_{\delta(s)-1}$, where $s_0 = s$, defined iteratively by

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

for $j = 1, 2, \dots, \delta(s)-1$. 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 duplication, 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 $a_{\delta(s)}, a_{\delta(s)-1}, \dots, a_1$, in this order.

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

$$S = \left(a_{\delta(s)}^+ \left(a_{\delta(s)-1}^+ \left(\dots \left(a_2^+ \left(a_1^+ \right)^+ \right) \right)^+ \right)^+ \right)^+.$$

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

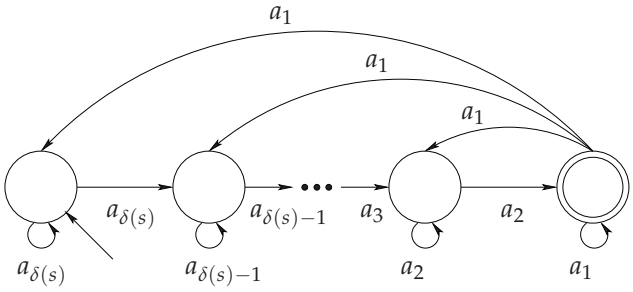

Figure 1. The finite-state automaton accepting the regular language used in the proof of Theorem 4.

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 [5], [10]). By Perron-Frobenius theory,

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

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

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

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

Proof: By applying Theorem 4 we get $\mathsf{cap}(S^{\mathrm{tan}}_{\geqslant 1}) \geqslant 1$. We also have the trivial upper bound $\mathsf{cap}(S^{\mathrm{tan}}_{\geqslant 1}) \leqslant \log_2 |\Sigma| = 1$, which completes the proof. ∎

For $S^{\mathrm{tan}}_{\geqslant k}$ and general k, we claim a weaker result, that is provided in the following theorem, stated without proof.

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 $\mathsf{cap}(S^{\mathrm{tan}}_{\geqslant k}) \geqslant \log_2 r > 0$, where r is the largest root of the polynomial*

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

V. Reversed Tandem Duplication

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

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

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 $s \in \Sigma^k$ with $s \neq s^R$. Then $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) \geqslant 1/k$.*

Proof: By repeatedly applying duplication 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, except the first one, (resp. any run of s^R) we can apply duplication to the last block of the previous run, which is an s^R block (resp. s). Thus, the regular language $S = ss^R \{s, s^R\}^*$, satisfies $S \subseteq S^{\mathrm{rt}}_k(s)$ and thus $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) \geqslant \mathsf{cap}(S)$. Furthermore, since $s \neq s^R$, we see that $\mathsf{cap}(S) = 1/k$. ∎

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

The following theorem states that the capacity of reversed tandem duplication 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: (Sketch). 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 only prove this for $|s| = k$. Denote $s = s_1 s_2 \dots s_k$, with $s_i \in \Sigma$, and let $t = s_2 s_3 \cdots s_k s_k$. It is not difficult to see that since $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) = 0$, we have $\mathsf{cap}(S^{\mathrm{rt}}_k(t)) = 0$. Hence, applying Lemma 7 to s and t implies that $s = s^R$ and $t = t^R$. From these, it can be shown that $\delta(s) = 1$. ∎

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 $S^{\mathrm{rt}}_k(s)$. This is formalized in the following lemma.

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 $S^{\mathrm{rt}}_k(y)$.*

Due to lack of space, we omit the proof of the lemma.

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 duplications, 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, $\mathsf{cap}(S^{\mathrm{rt}}_k(s)) = \mathsf{cap}(S^{\mathrm{rt}}_k(t))$. ∎

VI. Duplication with a Gap

Consider the duplication-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 $s = s_1 s_2 \dots$, with $s_i \in \Sigma$, we conveniently denote the substring starting at position i and of length k as $s_{i,k} = s_i s_{i+1} \dots s_{i+k-1}$. 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.

The following lemma, presented without proof, is useful for characterizing the set of sequences s with $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(s)) > 0$.

Lemma 11. *For all $s \in \Sigma^*$ such that $|s| \geqslant k + k'$, we have*

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

The next corollary is an immediate result of the lemma.

Corollary 12. *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*

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

This corollary is used in the following theorem.

Theorem 13. *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: (Sketch). 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. Here we only consider the case of $k > k'$. First, assume $s = x_1 \dots x_k y_1 \dots y_{k'}$, with $x_i, y_i \in \Sigma$, has length $k + k'$. Denote $k'' = k - k'$. We show that s is periodic with period $\gcd(k, k')$. From Corollary 12, 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 $T^{\mathrm{gap}}_{0,k,k'}$ to s we obtain

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

Consider 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 $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(t')) = 0$. By applying Corollary 12 to t', we get $x_1 \dots x_{k''} x_1 \dots x_{k'} = x_1 \dots x_{k'} x_1 \dots x_{k''}$, 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 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 at least 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 find a strict upper bound on $\mathsf{cap}(S^{\mathrm{gap}}_{k,k'}(s))$. For a sequence $x \in \Sigma^*$ and two symbols $a, b \in R(x)$, let

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

be the set of the differences of positions of a and b in x. Furthermore, let $\rho_{x,\ell}(a, b) = \{(j \bmod \ell) \mid j \in \Delta_x(a, b)\}$.

Lemma 14. *Let Σ be some finite alphabet, $d > 0$ an integer, and $D \subset \{0, 1, \dots, d-1\}$ 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 $\mathsf{cap}(S) < \log_2 |\Sigma|$.*

We omit the proof of Lemma 14 and only mention that it relies on the Perron-Frobenius theory. Using Lemma 14 we obtain the following theorem.

Theorem 15. *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 14. ∎

Our last result is the following theorem, which we state without proof.

Theorem 16. *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 $\delta(s)$.*

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] F. Farnoud, M. Schwartz, and J. Bruck, "The capacity of string-replication systems," *arXiv preprint: http://arxiv.org/abs/1401.4634*, 2014.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.