The Capacity of Some Pólya String Models

Ohad Elishco*, Farzad Farnoud (Hassanzadeh)†, Moshe Schwartz*, and Jehoshua Bruck†
*Electrical and Computer Engineering, Ben-Gurion University of the Negev,
Beer Sheva 8410501, Israel, {ohadeli@,schwartz@ee.}bgu.ac.il
†Electrical Engineering, California Institute of Technology
Pasadena, CA 91125, U.S.A., {farnoud,bruck}@caltech.edu

***Abstract*—We study random string-duplication systems, called Pólya string models, motivated by certain random mutation processes in the genome of living organisms. Unlike previous works that study the combinatorial capacity of string-duplication systems, or peripheral properties such as symbol frequency, this work provides exact capacity or bounds on it, for several probabilistic models. In particular, we give the exact capacity of the random tandem-duplication system, and the end-duplication system, and bound the capacity of the complement tandem-duplication system. Interesting connections are drawn between the former and the beta distribution common to population genetics, as well as between the latter system and signatures of random permutations.**

I. Introduction

Several mutation processes are known, which affect the genetic information stored in the DNA. Among these are transposon-driven repeats [5] and tandem repeats which are believed to be caused by slipped-strand mispairings [9]. In essence, these mutation processes take a substring of the DNA and insert a copy of it somewhere else (in the former case), or next to the original copy (in the latter). In human DNA, it is known that its majority consists of repeated sequences [5]. Moreover, certain repeats cause important phenomena such as chromosome fragility, expansion diseases, silencing genes [10], and rapid morphological variation [3].

A formal mathematical model for studying these kinds of mutation processes is the notion of *string-duplication systems*. In such systems, a seed string (or strings) evolves over time by successive applications of mutating functions. For example, functions taking a substring of a string and copying it next to itself model mutation by tandem duplication. These string-duplication systems were studied in the context of formal languages (e.g., [6]) in an effort to place the resulting sets of mutated sequences within Chomsky's hierarchy of formal languages, as well as derive closure properties.

Another approach, from a coding-theoretic perspective, attempted to find properties of string-duplication systems such as capacity and diversity [1], [4]. Using various techniques, mainly borrowed from constrained-coding theory, bounds were derived on the number of strings that are attainable via the studied mutation processes, given the original seed string. These were used to obtain either exact expressions or bounds on the *combinatorial capacity* of the string-duplication systems. The main drawback of this approach, however, is that in the expression for the combinatorial capacity, all attainable strings are considered equally likely. Clearly, there is a gap between this model and real-world mutation processes.

To reduce this gap, a probabilistic model was studied in [2]. This model is not concerned with which strings are possible, but rather with which strings are *probable*. With appropriate distributions applied to the choice of the mutated point, its length, and its final position, we obtain an induced distribution on resulting strings. However, [2] was not able to provide any exact capacity calculation nor bounds, and managed to study only peripheral properties of the resulting string distributions, namely the frequencies of symbols and substrings.

Thus, the goal of this paper is to find the exact capacity of probabilistic string-duplication systems, or bound it. As we later see, even for very modest parameters this problem is extremely challenging. The main contributions of this paper are an exact expression for the tandem-duplication system and the end-duplication system, as well as bounds on the capacity of the complement-tandem duplication system. In all cases we study duplication of length 1 only.

An important tool, widely used in the study of genetic drift in population genetics, is a *Pólya urn model*. It consists of an urn with balls of two different colors. In each step a ball is randomly chosen and returned to the urn along with k new balls of the same color [8]. There are many extensions of this model, where after each draw a set of balls, whose number and composition depends on the color of the drawn ball, are put into the urn. However, in these models there is no structure on the balls in the urn and only the number of balls of each color matters. Thus, these models fail to apply to strings.

We therefore suggest extensions of the Pólya urn models to what we call *Pólya string models*, in which the balls form a string, which may be circular or linear, similar to bases of a DNA molecule. The draw in this model typically involves choosing a random position (or equivalently a ball) in this string where a modification to the string–the mutation–occurs. In this paper, we focus on models in which after the draw, a sequence of balls is inserted to the string whose composition and the position it is inserted in depends on the local properties of the string around the chosen position.

The paper is organized as follows. In Section II we fix our notation and definitions that are used throughout the paper. In Section III we calculate the exact capacity of the tandem-duplication and end-duplication systems. In Section IV we bound the capacity of the complement tandem-duplication system. We conclude in Section V by providing some insight and comparisons with the combinatorial capacity and Pólya

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

urn models.

II. Preliminaries

Let $\Sigma = \{0,1\}$ be the binary alphabet. While the results we present have a greater generality, for the sake of simplicity of presentation we restrict ourselves to the binary case only. We use the notation common to formal languages to describe strings over Σ. The set of length-n strings over Σ is denoted by Σ^n. We let Σ^* denote the set of all finite-length strings over Σ. The unique empty string is denoted by ε. The set of all finite-length non-empty strings is denoted by $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$.

Let $\omega \in \Sigma^*$ be a string. We use $|\omega|$ to denote the length of ω. Obviously, $|\varepsilon| = 0$. If $\omega' \in \Sigma^*$, the concatenation of ω and ω' is denoted $\omega\omega'$. The number of occurrences of a symbol $a \in \Sigma$ in the string ω is denoted by $|\omega|_a$. For a natural number $n \in \mathbb{N}$ we use $[n]$ to denote the set $[n] = \{1, 2, \ldots, n\}$.

The Pólya string model may be quite generally defined. Intuitively, the model takes a starting string, and in a sequence of steps, mutates it over time. A formal definition follows:

Definition 1. *A Pólya string model is defined by $S = (\Sigma, \sigma, T)$, where Σ is a finite alphabet, $S(0) = \sigma \in \Sigma^+$ is a* seed string, *and $T : \Sigma^* \times \mathbb{N} \to \Sigma^*$ is a* duplication rule. *The string model is the following discrete-time random process: For all $i \in \mathbb{N}$ we let L_i be a random integer chosen uniformly from $[|S(i-1)|]$, i.e., an integer between 1 and the length of the string $S(i-1)$. We then set $S(i) = T(S(i-1), L_i)$.*

While the definition is given in terms of duplication mutations, it can be naturally extended to other types of mutations and random sequence editing.

Several rule choices parallel the combinatorial (deterministic) systems studied in [1], and are special cases of the general stochastic systems studied in [2]. In particular, we define the tandem duplication rule as

$$T^{\tan}(\omega a \omega', i) = \omega a a \omega',$$

where $\omega, \omega' \in \Sigma^*$, $|\omega| = i-1$, and $a \in \Sigma$. Intuitively, the tandem duplication rule takes the ith symbol of a given string and duplicates it next to the original letter. In a similar fashion we define the end-duplication and complement-tandem duplication rules as

$$T^{\mathrm{end}}(\omega a \omega', i) = \omega a \omega' a, \qquad T^{\overline{\tan}}(\omega a \omega', i) = \omega a \bar{a} \omega',$$

where $\omega, \omega' \in \Sigma^*$, $|\omega| = i-1$, $a \in \Sigma$, and $\bar{a}$ denotes the binary complement bit to a. By plugging in the appropriate duplication rule, we define the Pólya string systems $S^{\tan}$, S^{end}, and $S^{\overline{\tan}}$.

Given a Pólya string system S, the set of choices leading from $S(0)$ to $S(n)$ is denoted by $\mathcal{H}(n)$ and is referred to as the *history* of the sequence. The *capacity* of the process S is defined as

$$\mathsf{cap}(S) = \limsup_{n\to\infty} \frac{1}{n} H(S(n)),$$

where H is the entropy function (all logs are base 2),

$$H(S(n)) = -\sum_{\omega\in\Sigma^*} \Pr(S(n) = \omega) \log_2 \Pr(S(n) = \omega).$$

In a sense, the capacity quantifies the diversity that can be generated by the process. It also determines the smallest rate at which each symbol can be compressed. Furthermore, since $H(S(n)|\mathcal{H}(n)) = 0$,

$$\mathsf{cap}(S) = \limsup_{n\to\infty} \frac{1}{n} I(S(n); \mathcal{H}(n)),$$

where I denotes mutual information. Thus $\mathsf{cap}(S)$ can be viewed as the capacity the channel that transforms histories to sequences and can be used to derive rate-distortion results on estimating the history $\mathcal{H}(n)$ using the sequence $S(n)$.

III. Tandem and End Duplication

This section is dedicated to the study of the capacity of tandem and end duplication Pólya string models. In particular, this highlights the difference between an urn model and a string model. As we shall see, the capacity differs between some of the cases.

We start by stating the common points between the Pólya string models. We fix the binary alphabet $\Sigma = \{0,1\}$, and a starting string $S(0) = \sigma \in \Sigma^*$. Let us denote $|\sigma|_0 = t_0$ and $|\sigma|_1 = t_1$.

The random process repeatedly draws a position in uniform (independently of previous draws) from $S(i)$, $i = 0, 1, 2, \ldots$ and duplicates the bit in that position. Let us record the values of the chosen bits in each round as $\omega = b_0, b_1, b_2, \ldots$. The end result, after $n - t_0 - t_1$ rounds, is a binary string $S(n - t_0 - t_1)$ of length n. Let us further denote $|\omega|_0 = k_0$ and $|\omega|_1 = k_1$. Thus, $t_0 + t_1 + k_0 + k_1 = n$.

It is an easy exercise to find that the probability of recording a specific $\omega \in \Sigma^*$ depends only on t_0, t_1, k_0, and k_1, and does not depend on the order of bits in ω. In particular,

$$\Pr(\omega) = \frac{(t_0 + t_1 - 1)!(t_0 + k_0 - 1)!(t_1 + k_1 - 1)!}{(t_0 - 1)!(t_1 - 1)!(n-1)!}. \quad (1)$$

We now specialize our treatment depending on the duplication rule that is used.

Theorem 2. *For tandem duplication,* $\mathsf{cap}(S^{\tan}) = 0$.

Proof: We use a crude counting argument. Consider the initial string $S(0)$, and denote the number of runs in it by r. Obviously any tandem duplication operation extends existing runs and never creates new runs. Thus, it may be viewed as an action of throwing $n - t_0 - t_1$ balls into r bins. The total number of resulting strings (regardless of probability) is given exactly by $\binom{n-t_0-t_1+r-1}{r-1} \leqslant n^{r-1}$. Maximum entropy will be attained by a uniform distribution over those strings, and even in that case we get

$$\mathsf{cap}(S^{\tan}) \leqslant \limsup_{n\to\infty} \frac{1}{n} \log n^{r-1} = 0.$$

A lower bound of 0 is trivial since we have at least one string possible of each length $n \geqslant t_0 + t_1$. ■

The case of end duplication behaves quite differently.

Theorem 3. *Let $S(0) \in \Sigma^*$ be a seed string with $|S(0)|_0 = t_0$ and $|S(0)|_1 = t_1$. Then the capacity of the end-duplication Pólya string model with $S(0)$ is given by*

$$\mathsf{cap}(S^{\text{end}}(S(0))) = \int_0^1 \beta(p; t_0, t_1) H_2(p) dp,$$

where $H_2(\cdot)$ is the binary entropy function, and

$$\beta(p; t_0, t_1) = \frac{(t_0 + t_1 - 1)!}{(t_0 - 1)!(t_1 - 1)!} p^{t_0 - 1}(1 - p)^{t_1 - 1},$$

is the pdf for the $\mathrm{Beta}(t_0, t_1)$ *distribution.*

Proof: Consider the setting discussed above, in which we record the drawn bits $\omega = b_0, b_1, \ldots, b_{n-1-t_0-t_1}$. In the end-duplication case, the resulting string $S(n - t_0 - t_1) = S(0)\omega$ is simply the concatenation of ω to the seed string $S(0)$. Again, we denote by k_0 the number of 0's in ω, and by k_1 the number of 1's in ω.

The probability of drawing ω will be denoted by $\Pr(\omega)$, whereas, the probability of drawing any ω with k_0 0's will be denoted by $\Pr(k_0)$. By our previous discussion, all draws ω with the same number of 0's have the same probability, i.e.,

$$\Pr(k_0) = \binom{k_0 + k_1}{k_0} \Pr(\omega).$$

The capacity is now given by,

$$\begin{aligned}
&\mathsf{cap}(S^{\text{end}}(S(0))) \\
&\quad = \limsup_{n\to\infty} \frac{1}{n - t_0 - t_1} H(S(n - t_0 - t_1)) \\
&\quad = \limsup_{n\to\infty} \frac{1}{n} \sum_{k_0=0}^{n-t_0-t_1} - \left(\Pr(k_0) \log \Pr(\omega|\, |\omega|_0 = k_0)\right),
\end{aligned}$$

where in the last equality, $\Pr(\omega|\, |\omega|_0 = k_0)$ denotes the probability of a fixed ω with k_0 0's.

Plugging in the expression from (1) we get

$$\begin{aligned}
&\frac{1}{n} \sum_{k_0=0}^{n-t_0-t_1} - \left(\Pr(k_0) \log \Pr(\omega|\, |\omega|_0 = k_0)\right) \\
&\quad = -\frac{1}{n} \sum_{k_0=0}^{n-t_0-t_1} \Pr(k_0) \log \frac{(t_0 + k_0 - 1)!(t_1 + k_1 - 1)!}{(n-1)!} \\
&\qquad - \frac{1}{n} \log \frac{(t_0 + t_1 - 1)!}{(t_0 - 1)!(t_1 - 1)!},
\end{aligned}$$

and we note the last term is $o(1)$, i.e., it vanishes when n grows. We also have

$$\frac{(t_0 + k_0 - 1)!(t_1 + k_1 - 1)!}{(n-1)!} = \frac{1}{n-1} \binom{n-2}{t_0 + k_0 - 1}^{-1}.$$

Thus,

$$\begin{aligned}
&\frac{1}{n} \sum_{k_0=0}^{n-t_0-t_1} - \left(\Pr(k_0) \log \Pr(\omega|\, |\omega|_0 = k_0)\right) \\
&\quad = \frac{1}{n} \sum_{k_0=0}^{n-t_0-t_1} \Pr(k_0) \log \binom{n-2}{t_0 + k_0 - 1} + o(1).
\end{aligned}$$

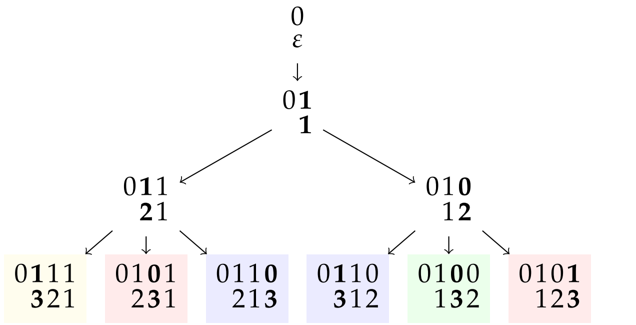

Figure 1. The tree of sequences that can be obtained starting from 0 using the complement tandem-duplication rule for $n \leqslant 3$. The first line in each node is the sequence and the second line is its history permutation.

Additionally, it is well known (e.g., see [8, Ch. 3]) that as $n \to \infty$, we have $\Pr(k_0/n \leqslant p) \to \int_0^p \beta(u; t_0, t_1) du$. Finally,

$$\begin{aligned}
&\mathsf{cap}(S^{\text{end}}(S(0))) \\
&\quad = \limsup_{n\to\infty} \frac{1}{n} \sum_{k_0=0}^{n-t_0-t_1} \Pr(k_0) \log \binom{n-2}{t_0 + k_0 - 1} \\
&\quad = \int_0^1 \beta(p) H_2(p) dp,
\end{aligned}$$

where we also used the well-known approximation for the binomial coefficient $\binom{n-2}{t_0+k_0-1} = 2^{nH_2(k_0/n)+o(n)}$ (e.g., see [7]). ■

IV. Complement Tandem Duplication

In this section, we consider the complement tandem-duplication Pólya string model, $S = S^{\overline{\tan}} = (\{0,1\}, \sigma, T^{\overline{\tan}})$. For simplicity, in what follows we assume that $S(0) = \sigma = 0$. Since we only have one choice, the string then becomes $S(1) = 01$. As an example, a possible history leading to $S(3) = 0110$ is

$$0 \to 0\mathbf{1} \to 01\mathbf{0} \to 0\mathbf{1}10, \tag{2}$$

where in each step the new symbol is in bold.

A history can be encoded as a permutation of length n, called its *history permutation*, as follows: Replace each 0 or 1 with the number of the turn in which they were added to the sequence. For example, the history given in (2) corresponds to the history permutation 312:

$$\begin{aligned}
0 &\to 0\mathbf{1} \to 01\mathbf{0} \to 0\mathbf{1}10, \\
\varepsilon &\to \ \ \mathbf{1} \to \ \ 1\mathbf{2} \to \ \ \mathbf{3}12.
\end{aligned}$$

Note that since 0 is always in the starting position, we drop it to obtain a permutation of $[n]$. It is clear that this provides us with a bijection between permutations of $[n]$ and a history resulting in a sequence $S(n) = 01s$, $s \in \{0,1\}^{n-1}$. This bijection will be useful in what follows.

The tree in Fig. 1 illustrates the history permutations and the sequences arising from them for $n \leqslant 3$. Since all histories are equally likely, all leaves at the same level in the tree are equally likely. Note however that not all sequences are equally likely as multiple histories may lead to the same sequence. For example, from Fig. 1, it is clear that $\Pr(S(3) = 0101) = 2\Pr(S(3) = 0100)$.

The following definitions will be useful. For $n \in \mathbb{N}$ let $\mathbb{S}_n$ denote the symmetric group of permutations over $[n]$. The ith element of $S(n)$, for $i \in [n+1]$, is denoted by $S_i(n)$. Furthermore, the substring $s_i s_{i+1} \cdots s_j$ of a sequence s is denoted by s_i^j. For $S(n)$, this notation becomes $S_i^j(n)$.

For a permutation $\pi \in \mathbb{S}_n$, define the signature $s \in \{0,1\}^{n-1}$ of π as

$$s_i = \begin{cases} 0, & \text{if } s_i > s_{i+1} \\ 1, & \text{if } s_i < s_{i+1} \end{cases}$$

for $i \in [n-1]$, i.e., ascents are marked by 1 and descents by 0.

The following theorem is useful in computing the capacity of the system.

Theorem 4. *The probability* $\Pr(S(n) = 01s)$, $s \in \{0,1\}^{n-1}$, *in* $S^{\overline{\tan}}$ *with* $S(0) = 0$, *is the same as the probability of getting the signature* s *when choosing a random permutation from* $\mathbb{S}_n$.

Proof: Let the set of history permutations that lead to $01s$ be denoted by Π_{01s}. Furthermore, let the set of permutations with signature $s \in \{0,1\}^{n-1}$ be denoted by Ψ_s. We claim that

$$|\Pi_{01s}| = |\Pi_{10s}| = |\Psi_s|, \tag{3}$$

for all $s \in \{0,1\}^*$. We show this by proving that the sizes of both sets satisfy the same recursion with the same initial values. The initial conditions for all recursions are $|\Pi_{01\varepsilon}| = |\Pi_{10\varepsilon}| = |\Psi_\varepsilon| = 1$, where ε is the empty string.

We start by providing two recursions for $|\Psi_s|$. For $v \in \{0,1\}^n$, let

$$\begin{aligned} T_v = \{i \in [n+1] : (&v_{i-1} = 1 \text{ or } i = 1) \\ &\text{and } (v_i = 0 \text{ or } i = n+1)\}, \\ U_v = \{i \in [n+1] : (&v_{i-1} = 0 \text{ or } i = 1) \\ &\text{and } (v_i = 1 \text{ or } i = n+1)\}, \end{aligned}$$

be the set of positions where 1 to 0 and 0 to 1 transitions occur (except at the boundaries). For example for $s = 0011010$, we have $T_s = \{1,5,7\}$ and $U_s = \{3,6,8\}$.

For $s \in \{0,1\}^n$, we can construct a permutation of $n+1$ elements with the signature s recursively by first determining the position of $n+1$. The set of valid positions for $n+1$ is precisely the set T_s. Suppose we place $n+1$ in position $i \in T_s$. We now need to construct two permutations with signatures $s_1 s_2 \cdots s_{i-2}$ and $s_i s_{i+1} \cdots s_n$ each with a subset of $[n]$. We can choose the set of elements for each of these two permutations in $\binom{n}{i-1}$ ways. Hence,

$$|\Psi_s| = \sum_{i \in T_s} \binom{n}{i-1} \left|\Psi_{s_1^{i-2}}\right| \left|\Psi_{s_{i+1}^n}\right|.$$

Similarly, by deciding where to place 1 (instead of $n+1$), we can show that

$$|\Psi_s| = \sum_{i \in U_s} \binom{n}{i-1} \left|\Psi_{s_1^{i-2}}\right| \left|\Psi_{s_{i+1}^n}\right|.$$

We now return to Π_{01s} and Π_{10s}. Note that (3) holds trivially if s is the empty string. Suppose (3) holds for all $s \in \{0,1\}^{n-1}$. Fix $s \in \{0,1\}^n$ and consider the sequence $01s$ as the result of the Pólya string model. In the permutations in Π_{01s}, the set of valid positions for 1 is precisely the set of positions in T_s. To see this note that in a permutation describing the history of $01s$, the element 1 can only correspond to the last element in a run of 1s in the string $01s$. Specifically, the element 1 can be placed in position 1 iff s starts with a 0 (since the bold 1 in $0\mathbf{1}s$ is the last 1 in a run); 1 can be placed in position $2 \leqslant i \leqslant n$ iff $s_{i-1}s_i = 10$; and finally, 1 can be placed in position $n+1$ iff $s_n = 1$ (again, the last 1 in a run of 1s).

Hence, we can construct these permutations recursively by first determining the position of 1 in them, and

$$\begin{aligned} |\Pi_{01s}| &= \sum_{i \in T_s} \binom{n}{i-1} \left|\Pi_{01s_1^{i-2}}\right| \left|\Pi_{10s_{i+1}^n}\right| \\ &= \sum_{i \in T_s} \binom{n}{i-1} \left|\Psi_{s_1^{i-2}}\right| \left|\Psi_{s_{i+1}^n}\right|. \end{aligned}$$

Similarly, for Π_{10s}, $s \in \{0,1\}^n$, the possible positions for 1 are precisely those in U_s as now 1 in the history permutation should correspond to the last 0 in a run of 0s in the string $10s$. So 1 can be placed in position 1 iff s starts with a 1; it can be placed in position $2 \leqslant i \leqslant n$ iff $s_{i-1}s_i = 01$; and finally it can be placed in position $n+1$ if $s_n = 0$. We thus have

$$\begin{aligned} |\Pi_{10s}| &= \sum_{i \in T_s} \binom{n}{i-1} \left|\Pi_{10s_1^{i-2}}\right| \left|\Pi_{01s_{i+1}^n}\right| \\ &= \sum_{i \in T_s} \binom{n}{i-1} \left|\Psi_{s_1^{i-2}}\right| \left|\Psi_{s_{i+1}^n}\right|. \end{aligned}$$

This completes the proof of (3) for all $s \in \{0,1\}^*$. ■

Define the process $\bar{S}$ as follows. Suppose we uniformly and independently choose random reals in $[0,1]$ denoted by $X_1, X_2, \ldots$. Let

$$\bar{S}_i = \begin{cases} 1, & \text{if } X_i < X_{i+1} \\ 0, & \text{if } X_i > X_{i+1} \end{cases} \tag{4}$$

for $i \in \mathbb{N}$. Note that the strings in S evolve by changing at a random position, but $\bar{S}$ can be viewed as evolving by changing at the end, and thus is easier to analyze. The key to our bounds on the capacity is drawing an equivalence between these two processes, which follows from Theorem 4: For any n and $s \in \{0,1\}^{n-1}$, we have

$$\Pr(S(n) = 01s) = \Pr(\bar{S}_1^{n-1} = s).$$

So,

$$\begin{aligned} \mathsf{cap}(S^{\overline{\tan}}) &= \limsup_{n\to\infty} \frac{1}{n} H(S(n)) = \limsup_{n\to\infty} \frac{1}{n} H\left(\bar{S}_1^{n-1}\right) \\ &= \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n-1} H\left(\bar{S}_i \middle| \bar{S}_1^{i-1}\right) \end{aligned} \tag{5}$$

Theorem 5. *For the complement tandem-duplication Pólya string model with seed string $S(0) = 0$, we have*

$$\frac{5\log e - 2}{6} \leqslant \mathsf{cap}(S^{\overline{\tan}}) \leqslant H_2\left(\frac{1}{3}\right).$$

Proof: Before proceeding with the proof, we show a simpler lower bound than the one given in the theorem. For $i \in \mathbb{N}$, since $\bar{S}_1^{i-1} \to X_i \to S_i$, i.e., they form a Markov chain, we have $H(\bar{S}_i|\bar{S}_1^{i-1}) \geqslant H(\bar{S}_i|X_i)$. Furthermore, $\Pr(\bar{S}_i = 0|X_i = x) = x$. Thus from (5) we find

$$\mathsf{cap}(S^{\overline{\tan}}) \geqslant H\left(\bar{S}_i|X_i\right) = \int_0^1 H_2(x)dx = \frac{\log e}{2} \geqslant 0.7213.$$

With the same approach we can prove the lower bound in the theorem. Note that $\bar{S}_1^{i-2} \to X_{i-1} \to \bar{S}_{i-1}^{i}$. So

$$\begin{aligned} H(\bar{S}_i|\bar{S}_1^{i-1}) &\geqslant H(\bar{S}_i|\bar{S}_{i-1}, X_{i-1}) \\ &= \int_0^1 xh_0(x)dx + \int_0^1 (1-x)\, h_1(x)dx, \end{aligned}$$

where $h_0(x) = H\left(\bar{S}_i|\bar{S}_{i-1} = 0, X_{i-1} = x\right)$ and $h_1(x) = H\left(\bar{S}_i|\bar{S}_{i-1} = 1, X_{i-1} = x\right)$. We have

$$\begin{aligned} h_0(x) &= H_2\left(\frac{1}{x}\int_0^x ydy\right) = H_2\left(\frac{x}{2}\right), \\ h_1(x) &= H_2\left(\frac{1}{1-x}\int_x^1 (1-y)\,dy\right) = H_2\left(\frac{1-x}{2}\right). \end{aligned}$$

Hence,

$$\begin{aligned} &H(\bar{S}_i|\bar{S}_1^{i-1}) = \int_0^1 xH_2\left(\frac{x}{2}\right)dx + \\ &\int_0^1 (1-x)\,H_2\left(\frac{1-x}{2}\right)dx = \frac{5\log e - 2}{6} \geqslant 0.8689. \end{aligned}$$

Now we turn to proving the upper bound. Note that

$$\begin{aligned} \mathsf{cap}(S^{\overline{\tan}}) &\leqslant \lim_{n\to\infty} H(\bar{S}_n|\bar{S}_{n-1}) = H(\bar{S}_2|\bar{S}_1) \\ &= \frac{1}{2}\left(H(\bar{S}_2|\bar{S}_1 = 0) + H(\bar{S}_2|\bar{S}_1 = 1)\right) \\ &= \frac{1}{2}\cdot 2\cdot H_2\left(\frac{1}{3}\right) \leqslant 0.9183, \end{aligned}$$

since by integrating over the values of X_1^3, we find

$$\Pr\left(\bar{S}_2 = 0|\bar{S}_1 = 0\right) = \frac{\int_0^1 dx_1 \int_0^{x_1} dx_2 \int_0^{x_2} dx_3}{\int_0^1 dx_1 \int_0^{x_1} dx_2} = \frac{1/6}{1/2} = \frac{1}{3}$$

as well as $\Pr\left(\bar{S}_2 = 1|\bar{S}_1 = 1\right) = \frac{1}{3}$. ■

Both methods used in the proof of the preceding theorem can be extended to obtain better bounds. We do this for the upper bound. We have

$$\mathsf{cap}(S^{\overline{\tan}}) \leqslant \lim_n H(\bar{S}_n|\bar{S}_{n-2}, \bar{S}_{n-1}) = H(\bar{S}_4|\bar{S}_2, \bar{S}_3)$$

Let $P_{ijk} = \Pr(\bar{S}_2 = i, \bar{S}_3 = j, \bar{S}_4 = k)$. By integration, we find $(P_{000}, P_{001}, \ldots, P_{111}) = \frac{1}{24}(1,3,5,3,3,5,3,1)$. Hence

$$\begin{aligned} H(\bar{S}_4|\bar{S}_2 = 0, \bar{S}_3 = 0) = H(\bar{S}_4|\bar{S}_2 = 1, \bar{S}_3 = 1) = H_2(2/8), \\ H(\bar{S}_4|\bar{S}_2 = 0, \bar{S}_3 = 1) = H(\bar{S}_4|\bar{S}_2 = 1, \bar{S}_3 = 0) = H_2(3/8). \end{aligned}$$

So $H(\bar{S}_4|\bar{S}_2, \bar{S}_3) = 2\cdot\frac{1}{6}H_2(2/8) + 2\cdot\frac{1}{3}H_2(3/8) \leqslant 0.9067$. With the same method, numerically, we can show that $\mathsf{cap}(S^{\overline{\tan}}) \leqslant 0.9045$. So,

$$0.8689 \leqslant \mathsf{cap}(S^{\overline{\tan}}) \leqslant 0.9045.$$

V. Conclusion

In this paper we defined and studied Pólya string models. The exact capacity of the tandem-duplication and end-duplication models was derived. In the case of complement tandem duplication we gave bounds on the capacity. We make several interesting observation. First, had we used a Pólya urn model instead of a string model, then no difference would have been observed between tandem and end duplication. Indeed, the distribution of 0's and 1's in both cases is the same. However, when considering the structure of a string, the difference between the two comes to light. Additionally, while the combinatorial capacity of end-duplication is known to be 1, in the probabilistic model it varies depending on the starting string. Similarly, for the complement tandem-duplication model, it is easy to show that the combinatorial capacity is 1, while the probabilistic capacity is bounded away from both 0 and 1. An obvious question that is still open, is to find solutions to the above-studied models when the duplication length is greater than 1.

References

[1] F. Farnoud, M. Schwartz, and J. Bruck, "The capacity of string-duplication systems," *IEEE Trans. Inform. Theory*, accepted.

[2] ——, "A stochastic model for genomic interspersed duplication," in *Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT2015), Hong Kong, China SAR*, Jun. 2015, pp. 1731–1735.

[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] S. Jain, F. Farnoud, and J. Bruck, "Capacity and expressiveness of genomic tandem duplication," in *Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT2015), Hong Kong, SAR China*, Jun. 2015, pp. 1946–1950.

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

[7] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. North-Holland, 1978.

[8] H. Mahmoud, *Pólya Urn Models*, 1st ed. Chapman & Hall/CRC, 2008.

[9] N. I. 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.

[10] 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.