Two-Pattern Strings III — Frequency of Occurrence and Substring Complexity

Frantisek Franek Jiandong Jiang W. F. Smyth *

Algorithms Research Group
Department of Computing & Software
McMaster University
Hamilton, Ontario
Canada L8S 4K1

June 27, 2005

Abstract

The two previous papers in this series introduced a class of infinite binary strings, called two-pattern strings, that constitute a significant generalization of, and include, the much-studied Sturmian strings. The class of two-pattern strings is a union of a sequence of increasing (with respect to inclusion) subclasses TP_λ of two-pattern strings of scope λ, $\lambda = 1, 2, \cdots$. Prefixes of two-pattern strings are interesting from the algorithmic point of view (their recognition, generation, and computation of repetitions and near-repetitions) and since they include prefixes of the Fibonnaci and the Sturmian strings, they merit investigation of how many finite two-pattern strings of a given size there are among all binary strings of the same length. In this paper we first consider the frequency $f_\lambda(n)$ of occurrence of two-pattern strings of length n and scope λ among all strings of length n on $\{a, b\}$: we show that $\lim_{n\to\infty} f_\lambda(n) = 0$, but that for strings of lengths $n \leq 2\lambda$, two-pattern strings of scope λ constitute more than one-quarter of all strings. Since the class of Sturmian strings is a subset of two-pattern

*also at School of Computing, Curtin University, Perth WA 6845, Australia, and Department of Computer Science, King's College London.

strings of scope 1, it was natural to focus the study of the substring complexity of two-pattern strings to those of scope 1. Though preserving the aperiodicity of the Sturmian strings, the generalization to two-pattern strings greatly relaxes the constrained substring complexity (the number of distinct substrings of the same length) of the Sturmian strings. We derive upper and lower bounds on $C_1(k)$ (the number of distinct substring of length k) of two-pattern strings of scope 1, and we show that it can be considerably greater than that of a Sturmian string. In fact, we describe circumstances in which $\lim_{k\to\infty}(C_1(k)-k) = \infty$.

1 Introduction

This paper is a sequel to [FLS03, FLS04] that we recommend to the attention of the reader. However, we make this paper self-contained by briefly reviewing the essential definitions already provided, particularly in [FLS03]. Terminology and notation generally follow [S03]. For the sake of simplicity, **string** refers to a finite binary string on the alphabet $\{a, b\}$; for infinite binary strings on $\{a, b\}$ we will use the explicit reference **infinite string**.

Suppose an integer $\lambda \geq 1$ is given (the ***scope***), together with nonempty strings $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\{a, b\}$ such that $|\boldsymbol{p}| \leq \lambda, |\boldsymbol{q}| \leq \lambda$. We call $\boldsymbol{p}$ and $\boldsymbol{q}$ ***patterns of scope*** λ, and we suppose that they are ***suitable*** (see below for details — roughly speaking, $\boldsymbol{p}$ and $\boldsymbol{q}$ are constrained to be dissimilar enough that they can be efficiently distinguished from each other). For any pair of positive integers i and j, $i \neq j$, consider the morphism that maps single letters into ***blocks***:

$$\sigma : \quad a \to \boldsymbol{p}^i\boldsymbol{q}, \ b \to \boldsymbol{p}^j\boldsymbol{q}. \tag{1}$$

We call σ an ***expansion of scope*** λ and denote it by the 4-tuple $[\boldsymbol{p}, \boldsymbol{q}, i, j]_\lambda$ (or just $[\boldsymbol{p}, \boldsymbol{q}, i, j]$ if the scope is clear from the context).

Of course an expansion can be applied to any (finite or infinite) string $\boldsymbol{x}$ oin $\{a, b\}$ by defining

$$\sigma(\boldsymbol{x}) = \sigma\big(\boldsymbol{x}[1]\big)\sigma\big(\boldsymbol{x}[2]\big)\cdots\sigma\big(\boldsymbol{x}[n]\big)\cdots,$$

and the composition of two expansions is equally well-defined:

$$(\sigma_2 \circ \sigma_1)(\boldsymbol{x}) = \sigma_2\big(\sigma_1(\boldsymbol{x})\big).$$

Suppose a finite sequence

$$\sigma_1, \sigma_2, \ldots, \sigma_k$$

of expansions of scope λ is given. Then the string

$$\boldsymbol{x} = (\sigma_k \circ \sigma_{k-1} \circ \cdots \circ \sigma_1)(a) \tag{2}$$

is called a ***complete two-pattern string*** of scope λ. (More generally, we call any substring of $\boldsymbol{x}$ a ***two-pattern string*** of scope λ.)

Conversely, if it is known that a (finite or infinite) string $\boldsymbol{x}$ is a concatenation of blocks $\boldsymbol{p}^i\boldsymbol{q}$ and $\boldsymbol{p}^j\boldsymbol{q}$, then a ***reduction*** ρ is well-defined on $\boldsymbol{x}$ by

$$\rho: \quad \boldsymbol{p}^i\boldsymbol{q} \to a, \ \boldsymbol{p}^j\boldsymbol{q} \to b, \tag{3}$$

and we say that $\boldsymbol{x}$ is ***reducible*** by ρ. An infinite string $\boldsymbol{x}$ is called a ***infinite complete two-pattern string*** of scope λ if and only if its every prefix is a prefix of a complete two-pattern string of scope λ.
Note: (a) if $\boldsymbol{x}$ is an infinite complete two-pattern string of scope λ, and σ is an expansion of scope λ, then $\sigma(\boldsymbol{x})$ is an infinite complete two-pattern strings of scope λ;
(b) if $\boldsymbol{x}$ is an infinite complete two-pattern string of scope λ, then there exists at least one reduction of $\boldsymbol{x}$, and for any reduction ρ of $\boldsymbol{x}$, $\rho(\boldsymbol{x})$ is an infinite complete two-pattern string of scope λ.
More generally, any suffix of a infinite complete two-pattern of scope λ is called an ***infinite two-pattern string*** of scope λ.

Observe that every complete two-pattern string of scope λ is a prefix of infinitely many infinite complete two-pattern strings of scope δ, for any $\delta \geq \lambda$.

In the case $\lambda = 1$, for any expansion, $\boldsymbol{p}$ and $\boldsymbol{q}$ must both be single-letter strings, and the suitability condition requires that $\boldsymbol{p} = a$, $\boldsymbol{q} = b$ (or of course *vice versa*). If the further restriction that $j = i \pm 1$ is applied in every expansion, then the corresponding infinite two-pattern strings are in fact Sturmian, and vice versa, every infinite Sturmian string is an infinite two-pattern string of scope 1.

In [FKS00] we showed how to recognize prefixes of infinite Sturmian strings in time proportional to their length, a result extended in [FLS03] to complete two-pattern strings. In [FLS04] we described an algorithm to compute all the repetitions and near repetitions in linear time for complete

two-pattern strings, again extending the same result on prefixes of infinite Sturmian strings [FKS00].

In order to make these results intelligible, we now keep our promise to define a suitable pair of patterns. To do this we first need to define the idea of "regularity".

Definition 1 *A string* $\boldsymbol{q}$ *is said to be* $\boldsymbol{p}$**-regular** *if and only if there exist (possibly empty) strings* $\boldsymbol{u}, \boldsymbol{v}$ *and integers* $k \geq 1, r \geq 0$ *such that*

$$\boldsymbol{q} = (\boldsymbol{w}\boldsymbol{p}^{h_1})(\boldsymbol{w}\boldsymbol{p}^{h_2}) \cdots (\boldsymbol{w}\boldsymbol{p}^{h_k})\boldsymbol{u},$$

where

- $\boldsymbol{w} = \boldsymbol{u}\boldsymbol{p}^r\boldsymbol{v}$ *(*$\boldsymbol{v}$ *empty if* $r = 0$*);*
- $\boldsymbol{p}$ *is neither a prefix nor a suffix of either* $\boldsymbol{u}$ *or* $\boldsymbol{v}$*:*
- *each* h_j*,* $j = 1, 2, \ldots, k$*, takes one of only two nonnegative integer values; that is,*
$$1 \leq \left|\{h_j : \ j \in 1..k\}\right| \leq 2.$$

Thus if $\boldsymbol{q}$ is $\boldsymbol{p}$-regular, it contains $k \geq 1$ occurrences of $\boldsymbol{w}$, hence at least kr occurrences of $\boldsymbol{p}$; furthermore, $\boldsymbol{q}$ has a prefix consisting of "almost regular" sections $\boldsymbol{w}\boldsymbol{p}^{h_j}$, where the j^{th} section contains $r+h_j$ occurrences of $\boldsymbol{p}$. Thus, in rough terms, if $\boldsymbol{q}$ is $\boldsymbol{p}$-regular, then it is "built from $\boldsymbol{p}$ in a very particular and regular way".

In the definition of suitability for the pair $(\boldsymbol{p}, \boldsymbol{q})$ it is required that $\boldsymbol{q}$ be *not* $\boldsymbol{p}$-regular, thus the more restrictive the definition of regularity is, the bigger the number of suitable pairs, and the bigger the class of two-pattern strings. For technical reasons we used in [FLS03] and [FLS04] a more relaxed definition of regularity, however here we present a paraphrase of the more restrictive definition stated at the end of [FLS03] to obatin as large class of two-pattern strings as possible.

Definition 2 *An ordered pair* $(\boldsymbol{p}, \boldsymbol{q})$ *of nonempty strings is said to be* **suitable** *if and only if*

- $\boldsymbol{p}$ *is* **primitive** (*that is, in our use of the term,* $\boldsymbol{p}$ *has no nonempty border*)*;*
- $\boldsymbol{p}$ *is not a suffix of* $\boldsymbol{q}$*;*

- $\boldsymbol{q}$ *is neither a prefix nor a suffix of* $\boldsymbol{p}$*;*
- $\boldsymbol{q}$ *is not* $\boldsymbol{p}$*-regular.*

In Section 2 we study the frequency of occurrence of complete two-pattern strings of length n among all strings of length n; then in Section 3 we go on to derive upper and lower bounds on the substring complexity of infinite two-pattern strings of scope 1. Finally, in Section 4, we discuss open problems.

2 Frequency of Occurrence

If $T_\lambda(n)$ is the number of complete two-pattern strings of scope λ and length n, then the ***frequency*** $f_\lambda(n)$ of such strings among all strings of length n on $\{a, b\}$ is defined by

$$f_\lambda(n) = T_\lambda(n)/2^n.$$

Our first result is unsurprising:

Theorem 1 *For any fixed* λ, $\lim_{n\to\infty} f_\lambda(n) = 0$.

Proof Observe that a complete two-pattern string of scope λ is a concatenation of blocks of two types only: $A = \boldsymbol{p}^i\boldsymbol{q}$, and $B = \boldsymbol{p}^j\boldsymbol{q}$. Since $|\boldsymbol{p}|, |\boldsymbol{q}| \le \lambda$, there are less than $2^{\lambda+1}$ of distinct $\boldsymbol{p}$'s and $\boldsymbol{q}$'s. Therefore there are less than $2^{2\lambda+2}$ of distinct suitable pairs ($\boldsymbol{p}$,$\boldsymbol{q}$). It follows that there are less than $n{\cdot}2^{2\lambda+2}$ of distinct A's and B's, therefore less than $n^2{\cdot}2^{4\lambda+4}$ of distinct pairs (A,B)'s. Since $|A|, |B| \ge 2$, the two blocks can be concatenated together in at most $2^{\frac{n}{2}}$ different ways. Thus, $T_\lambda(n) < 2^{\frac{n}{2}}{\cdot}n^2{\cdot}2^{4\lambda+4}$ for any n, and, consequently, $f_\lambda(n) = \frac{T_\lambda(n)}{2^n} \le \frac{2^{\frac{n}{2}}{\cdot}n^2{\cdot}2^{4\lambda+4}}{2^n} = \frac{n^2{\cdot}2^{4\lambda+4}}{2^{\frac{n}{2}}}$. Since λ is a fixed constant, $f_\lambda(n) \to 0$ when $n \to \infty$, as $\frac{n^2}{2^{\frac{n}{2}}} \to 0$. $\square$

More interesting, and more descriptive of complete two-pattern strings, are results describing their frequency of occurrence when λ is large relative to n. We consider the number of complete two-pattern strings $\boldsymbol{x} = \boldsymbol{pq}$, where $(\boldsymbol{p}, \boldsymbol{q})$ is a suitable pair satisfying $|\boldsymbol{p}| \le \lambda$, $|\boldsymbol{q}| \le \lambda$. In order for such strings to exist, we must have $\lambda \ge \lceil n/2 \rceil$, and in fact we suppose $\lambda = \lceil n/2 \rceil$ for the following discussion.

Let π_k denote the number of primitive binary strings of length k, and let $\phi_k = \pi_k/2^k$. It is well known [GO81] that ϕ_k is monotone decreasing and rapidly convergent to the constant $\phi = 0.26778684\cdots$.

For $n \geq 6$ and $\lambda = \lceil n/2 \rceil$, we consider $\boldsymbol{x} = \boldsymbol{pq}$ for n even and n odd:

- If n is even, then $n = 2\lambda$, $\lambda \geq 3$, and $|\boldsymbol{p}| = |\boldsymbol{q}| = \lambda$. Observing from Definition 2 that $\boldsymbol{x}$ will in this case certainly be a two-pattern string if $\boldsymbol{p}$ is primitive, and observing further that $\boldsymbol{q}$ may therefore be any string of length λ except $\boldsymbol{p}$, we write

$$\begin{aligned} f_\lambda(n) &\geq \pi_\lambda(2^\lambda - 1)/2^n \\ &= \phi_\lambda\left(1 - \frac{1}{2^\lambda}\right) \\ &> 7\phi/8. \end{aligned} \tag{4}$$

- If n is odd, then $n = 2\lambda - 1$ and $\lambda \geq 4$. For $|\boldsymbol{p}| = \lambda - 1, |\boldsymbol{q}| = \lambda$, Definition 2 tells us that for primitive $\boldsymbol{p}$, $\boldsymbol{q}$ can be any string of length λ that does not contain $\boldsymbol{p}$ as a substring. Thus

$$\begin{aligned} f_\lambda(n) &\geq \pi_{\lambda-1}(2^\lambda - 4)/2^n \\ &= \phi_{\lambda-1}\left(1 - \frac{1}{2^{\lambda-2}}\right) \\ &> 3\phi/4. \end{aligned} \tag{5}$$

  On the other hand, for $|\boldsymbol{p}| = \lambda, |\boldsymbol{q}| = \lambda - 1$, $\boldsymbol{q}$ can be any string of length $\lambda - 1$ that is not a substring of $\boldsymbol{p}$, and so

$$\begin{aligned} f_\lambda(n) &\geq \pi_\lambda(2^{\lambda-1} - 2)/2^n \\ &= \phi_\lambda\left(1 - \frac{1}{2^{\lambda-2}}\right). \end{aligned} \tag{6}$$

  Since unfortunately these two cases are not independent, (5) and (6) are not additive.

Using brute force, one can compute $f_{\lceil n/2 \rceil}(n)$ for small n:

n	$f_{\lceil n/2 \rceil}(n)$
2	1/2
3	1/2
4	1/2
5	5/8

From (4) and (6) we have then

Theorem 2 *For $n \geq 2$,*

$$\begin{aligned} f_{\lceil n/2 \rceil}(n) &> \phi\left(1-\frac{1}{2^{\lceil n/2 \rceil}}\right) > 7\phi/8, \ n \text{ even;} \\ &> \phi\left(1-\frac{1}{2^{\lceil n/2 \rceil - 2}}\right) > 3\phi/4, \ n \text{ odd.} \ \square \end{aligned}$$

Thus, for $\lceil n/2 \rceil \leq \lambda < n$, $f_\lambda(n) \geq f_{\lceil n/2 \rceil}(n)$, and so $f_\lambda(n)$ is bounded below by a quantity that is close to ϕ for both even and odd n. In other words, over these values of λ, $T_\lambda(n)$ has a lower bound of roughly $2^n/4$ — more than one-quarter of all sufficiently long strings($n \geq 14$, say) are in fact two-pattern strings of some scope λ.

3 Substring Complexity

The observations of Section 2 encourage us to consider the ***substring complexity*** values $C_\lambda(k)$ of infinite complete two-pattern strings of scope λ; that is, for every integer $k \geq 0$, the number of distinct substrings of length k that may occur in the string. A Sturmian string is usually defined [L02, p. 45] to be an aperiodic infinite string that achieves the least complexity $C(k) = k+1$ for all $k \geq 0$.

In this section we obtain upper and lower bounds on $C_1(k)$ for infinite complete two-pattern strings. In the introduction it was disucssed that if $\boldsymbol{x}$ is an infinite complete two-pattern string, then it is reducible and its reduction is again an infinite complete two-pattern string.

We are assuming to have an infinite complete two-pattern strings $\boldsymbol{x}$. Moreover, we are assuming to have a reduction $\rho = [\boldsymbol{p}, \boldsymbol{q}, i, j]$ of $\boldsymbol{x}$. To simplify the proofs, we assume from now on that $i < j$ — the complexity results are unaffected — and we also assume without loss of generality that $\boldsymbol{p} = a$, $\boldsymbol{q} = b$. Thus, $\rho = [a, b, i, j]$.

For every $k \in 0..i$, $\boldsymbol{x}$ has exactly $k+1$ distinct substrings of length k,

$$a^k, \ a^{k-m-1}ba^m, m = 0, 1, \ldots, k-1; \tag{7}$$

and furthermore exactly one of these substrings, $\boldsymbol{u} = a^k$, is a prefix of two distinct substrings of $\boldsymbol{x}$, $\boldsymbol{u}a$ and $\boldsymbol{u}b$, of length $k+1$. Thus for every $k \in 0..i$, $C_1(k+1) = C_1(k)+1 = k+2$, just as for the Sturmian strings.

This simple observation leads to a useful notion: we shall say that a finite substring $\boldsymbol{u}$ of $\boldsymbol{x}$ is ***prolific*** if and only if both $\boldsymbol{u}a$ and $\boldsymbol{u}b$ are also substrings

of $\boldsymbol{x}$. The use of prolific substrings of $\boldsymbol{x}$ follows from the simple fact that $C_\lambda(k+1) = (C_\lambda(k) - P_k) + 2P_k = C_\lambda(k) + P_k$, where P_k is the number of distinct prolific substrings of $\boldsymbol{x}$ of length k.

Observe now that, again as for the Sturmian strings, $\boldsymbol{x}$ cannot be periodic, since in the morphisms ρ^{-1} that can be thought of as constructing it, it is true that $\boldsymbol{p} \neq \boldsymbol{q}$ and $i \neq j$. Thus [L02, p. 22]

$$C_\lambda(k{+}1) \geq C_\lambda(k) + 1 \tag{8}$$

for all $k \geq 1$, and since every distinct substring of length k is a prefix of at least one distinct substring of length $k{+}1$, there must exist at least one prolific substring $\boldsymbol{u_k}$ of every length k.

In fact, even for $\lambda = 1$, there may be several prolific substrings $\boldsymbol{u_k}$ in $\boldsymbol{x}$; consider, for example,

$$\boldsymbol{x} = \cdots abaaaaaaaaaababaaaaaaaaaaababababababaaaaaaaaaaab\cdots, \tag{9}$$

reducible by $[a, b, 1, 10]$ to $\rho(\boldsymbol{x}) = \cdots ababaaaaab\cdots$, a string that is in turn reducible by $\rho' = [a, b, 1, 5]$. We find that for $k = 6$, there are actually three prolific substrings

$$\boldsymbol{u_6} = aaaaaa,\ aababa,\ bababa, \tag{10}$$

so that in this case $C_1(7) = C_1(6){+}3$. Indeed, there are three prolific substrings $\boldsymbol{u_k}$ for every $k \in 6..9$ and two for every $k \in 10..11$. We shall see below that this multiplicity of prolific substrings for various values of k occurs because the substring *ababa* is "special".

Reflecting on this example, we are led to

Lemma 3 *Suppose that an infinite complete two-pattern string $\boldsymbol{x}$ reducible by $\rho = [a, b, i, j]$, $i < j$, contains a nonempty prolific substring $\boldsymbol{u}$. Then*

(a) either $\boldsymbol{u} = a^k$ *for some* $k \in 1..j-1$ or $\boldsymbol{u} = \boldsymbol{v}ba^i$ *for some (possibly empty) string* $\boldsymbol{v}$*;*

(b) *every suffix of* $\boldsymbol{u}$ *is prolific;*

(c) *if* $\boldsymbol{u} = a^k b\boldsymbol{v}$ *for some* $k \in 0..i$*, then* $a^i b\boldsymbol{v}$ *is prolific;*

(d) *if* $\boldsymbol{u} = a^k b\boldsymbol{v}$ *for some* $k \in i{+}1..j$*, then* $a^i ba^j b\boldsymbol{v}$ *is prolific.* □

Before embarking on further discussion of prolific strings, we need to identify two particular types of substrings: if $\rho(\boldsymbol{x})$ is reducible by $[a, b, i', j']$, we say that

- $\boldsymbol{u} = (a^i b)^{j'} a^i$ is ***exceptional*** in $\boldsymbol{x}$;
- $\boldsymbol{u} = (a^i b)^{j'+1} \boldsymbol{w}$ for some substring $\boldsymbol{w}$ is ***left-extendible*** in $\boldsymbol{x}$.

It is clear that if $\boldsymbol{u}$ is exceptional, then it is prolific, since both $(a^i b)^{j'} a^i a$ (a substring of $(a^i b)^{j'}(a^j b)$) and $(a^i b)^{j'} a^i b$ (a substring of $(a^i b)^{j'}(a^i b)^{j'}$) exist in $\boldsymbol{x}$. In particular, in the example (9) the exceptional string $(ab)^5 a$ is prolific — both $(ab)^5 a^2$ and $(ab)(ab)^5$ occur in $\boldsymbol{x}$. But note that these two occurrences are quite different: $(ab)^5 a^2$ can only occur preceded by b, while $(ab)(ab)^5$ must be preceded by a. Thus, in this case, the fact that $(ab)^5 a$ is prolific does *not* imply that

$$\rho\big((ab)^5\big) = a^5$$

is prolific in $\rho(\boldsymbol{x})$. As we shall soon discover, this circumstance is truly "exceptional".

The situation is different if the substring $\boldsymbol{u}$ is left-extendible: in this case, $\boldsymbol{u}$ is not necessarily prolific. However, the following result is easy to prove:

Lemma 4 *Suppose a string $\boldsymbol{u}$ is left-extendible in an infinite complete two-pattern string $\boldsymbol{x}$ of scope 1 reducible by $\rho = [a, b, i, j]$. Then $\boldsymbol{u}$ is prolific in $\boldsymbol{x}$ if and only if $a^i b a^{j-i} \boldsymbol{u}$ is prolific in $\boldsymbol{x}$.* □

In our following discussion of prolific strings, we shall exclude prolific substrings $\boldsymbol{u} = (a^i b)^{j'+1} \boldsymbol{w}$ that are left-extendible; Lemma 4 tells us that this exclusion is unimportant, since $\boldsymbol{u}$ is just a suffix of the prolific string

$$(a^i b)(a^j b)(a^i b)^{j'} \boldsymbol{w}.$$

This leads us to the notion of standard form:

If any substring $\boldsymbol{u}$ of an infinite complete two-pattern string $\boldsymbol{x}$ has prefix $a^i b$ and suffix $b a^i$ and is neither exceptional nor left-extendible, we say that $\boldsymbol{u}$ is in ***standard form****.*

Observe that by Lemma 3(c)-(d), every prolific substring of $\boldsymbol{x}$ that contains b is a suffix of a prolific substring of $\boldsymbol{x}$ in standard form.

Theorem 5 *Suppose that an infinite complete two-pattern string* $\boldsymbol{x}$ *is reducible by* $\rho = [a, b, i, j]$, $i < j$. *Let* $\boldsymbol{u} = \boldsymbol{v}a^i$ *be a substring of* $\boldsymbol{x}$ *in standard form. Then*

$$\boldsymbol{u} \text{ prolific in } \boldsymbol{x} \iff \rho(\boldsymbol{v}) \text{ prolific in } \rho(\boldsymbol{x}).$$

Proof If $\boldsymbol{u}$ is prolific in $\boldsymbol{x}$, both $\boldsymbol{u}a$ and $\boldsymbol{u}b$ occur in $\boldsymbol{x}$. Therefore, since $\boldsymbol{v}$ has prefix a^ib and suffix b, and since $\boldsymbol{u}$ is neither exceptional nor left-extendible, $\rho(\boldsymbol{v})$ is well-defined and both

$$\rho(\boldsymbol{v})b \quad \text{(corresponding to } \boldsymbol{u}a\text{)}$$

and

$$\rho(\boldsymbol{v})a \quad \text{(corresponding to } \boldsymbol{u}b\text{)}$$

occur in $\rho(\boldsymbol{x})$.

Conversely, if $\rho(\boldsymbol{v})$ is prolific in $\rho(\boldsymbol{x})$, both $\rho(\boldsymbol{v})a$ and $\rho(\boldsymbol{v})b$ must occur in $\rho(\boldsymbol{x})$. Hence for $\sigma = \rho^{-1}$, both

$$\boldsymbol{v}\sigma(a) = \boldsymbol{v}a^ib \quad \text{(corresponding to } \rho(\boldsymbol{v})a\text{)}$$

and

$$\boldsymbol{v}\sigma(b) = \boldsymbol{v}a^jb \quad \text{(corresponding to } \rho(\boldsymbol{v})b\text{)}$$

occur in $\boldsymbol{x}$. Thus $\boldsymbol{u} = \boldsymbol{v}a^i$ is prolific in $\boldsymbol{x}$. □

To better understand the meaning of this result, consider a prolific string $\boldsymbol{u} = a^ib\boldsymbol{v}a^i$ in standard form in an infinite complete two-pattern string $\boldsymbol{x}$ of scope 1 reducible by $\rho = [a, b, i, j]$, $i < j$. Let us call $a^ib\boldsymbol{v}$ the ***kernel*** of $\boldsymbol{u}$. Observe that by Theorem 5,

$$\rho(a^ib\boldsymbol{v}) = a\rho(\boldsymbol{v})$$

is prolific in $\rho(\boldsymbol{x})$, where $\rho(\boldsymbol{x})$ is reducible by $[a, b, i', j']$, $i' < j'$. Thus by Lemma 3(a), $a\rho(\boldsymbol{v})$ *either* takes the value a^k for some $k \in 1..j'-1$ *or* has suffix $a^{i'}$ and prefix a^kb for some $k \geq 1$. Supposing $a\rho(\boldsymbol{v}) \neq a^k$, we may *either* replace the prefix a^k in $a\rho(\boldsymbol{v})$ by $a^{i'}$ (if $a\rho(\boldsymbol{v})$ is not left-extendible) or remove the prefix $a^kb = a^{i'}b$ (if it is). Either way we get a prolific string

$$\boldsymbol{u'} = a^{i'}b\boldsymbol{v'}a^{i'}$$

in standard form with kernel $a^{i'}b\boldsymbol{v'}$. Note that

$$\begin{aligned} |a^{i'}b\boldsymbol{v'}| &= |a\rho(\boldsymbol{v})|+(i'-k)-i' \\ &\leq |a\rho(\boldsymbol{v})|-1 \\ &= |\rho(\boldsymbol{v})| \\ &\leq |\boldsymbol{v}|/2 \\ &= |a^ib\boldsymbol{v}|/2-(i+1)/2. \end{aligned}$$

Thus the kernel in $\rho(\boldsymbol{x})$ is less than half the length of the corresponding kernel in $\boldsymbol{x}$, and we have proved

Theorem 6 *Let $\boldsymbol{x}$ be an infinite complete two-pattern string of scope 1, reducible by $[a,b,i,j]$, and let $\boldsymbol{u}$ be a prolific substring of $\boldsymbol{x}$ in standard form. At most $r=\log_2(|\boldsymbol{u}|-i)$ reductions transform the kernel of $\boldsymbol{u}$ into a prolific substring a^k of the r^{th} reduction of $\boldsymbol{x}$.*

As we have seen, it may happen that a prolific string terminates, in the sense that it can no longer be extended to the left. For example, in (9), the substring

$$\boldsymbol{u}=(ab)^5a$$

of length 11 is prolific, with a kernel that transforms into a^4, a prolific substring of $\rho(\boldsymbol{x})$ that is reducible by $[a,b,1,5]$. However, there is no prolific substring $a\boldsymbol{u}$ or $b\boldsymbol{u}$.

On the other hand, prolific substrings (such as $ababa^{10}baba$ in (9)) may sometimes be indefinitely extendible to the left to form longer prolific substrings — this must always be true, for instance, in the Sturmian case, where there is only one prolific substring for each length k.

We can use Theorems 5 and 6 to generalize these observations and to extablish bounds on $C_1(k)$. Let us suppose that an infinite complete two-pattern string $\boldsymbol{x}$ of scope 1 is reducible by $\rho=[a,b,i,j]$, $i<j$, and $\rho(\boldsymbol{x})$ is reducible by $\rho'=[a,b,i',j']$, $i'<j'$. Then we may classify the prolific substrings of $\boldsymbol{x}$ in the range $0..j-1$ as follows:

(C1) For $k\in 0..j-1$, a^k is prolific. If $i=j-1$, this range reduces to $0..i$.

(C2) Consider $\boldsymbol{u}=(a^ib)^{j'}a^i=\boldsymbol{v}ba^i$ and observe that for every suffix $\boldsymbol{v'}$ of $\boldsymbol{v}$, $\boldsymbol{v'}ba^i$ is prolific in $\boldsymbol{x}$. But $b\boldsymbol{u}$ is not prolific, since $b\boldsymbol{u}b$ cannot occur in $\boldsymbol{x}$, while $a\boldsymbol{u}$ is prolific if and only if $i'=j'-1$. Thus the substrings in the sequence $ba^i..\boldsymbol{v}ba^i$ are all prolific for $k\in i+1..i+j'(i+1)$, and the sequence can be extended to $k=i+j'(i+1)+1$ if and only if $i'=j'-1$.

Note also that if both $i = j-1$ and $i' = j'-1$, there will be exactly one prolific string for every $k \in 0..i+j'j$, while for $i < j-1$, there will in view of (C1) be at least two prolific strings for every

$$k \in i+1..\min\{j-1, i+j'(i+1)\}.$$

(C3) Consider the substring $\boldsymbol{t} = a^{i+1}b(a^ib)^{i'}a^i$ of length $k = i+(i'+1)(i+1)+1$. This substring is prolific and indeed, for $i' = j'-1$, $\boldsymbol{t} = a\boldsymbol{u}$, while otherwise $|\boldsymbol{t}| < |\boldsymbol{u}|$. In fact, by Lemma 3(c)-(d), $a^ib(a^ib)^{i'}a^{j-i-1}\boldsymbol{t}$ of length $j+2(i'+1)(i+1)$ is also prolific. Thus if both $i = j-1$ and $i' = j'-1$, there is exactly one prolific substring for each $k \in 0..i+2j'j$; on the other hand, if both $i < j-1$ and $i' < j'-1$, then in view of (C1) and (C2) there will be three prolific substrings for every $k \in i+(i'+1)(i+1)+1..j-1$, a range that is nonempty whenever $i+(i'+1)(i+1) < j-1$.

We observe that the cases (C1)-(C3) exhaust all the possibilities for the range $0..j-1$: in the Sturmian case, both $i = j-1$ and $i' = j'-1$, so that as expected exactly one prolific string occurs for each k; while if both $i < j-1$ and $i' < j'-1$, there may be as many as three prolific strings for certain values of k. We observe further that the same result is true for any reduction of $\boldsymbol{x}$; since by Theorem 5 there must exist in $\boldsymbol{x}$ a corresponding range determined by the inverse expansions, we see that there may be ranges of values of k in $\boldsymbol{x}$ for which there exist three prolific strings. More than three is not possible, because the range must after a finite number of reductions reduce to $0..j-1$. Thus for every $k \geq i+1$,

$$C_1(k)+1 \leq C_1(k+1) \leq C_1(k)+3,$$

and with a little calculation we can establish

Theorem 7 *Let $\boldsymbol{x}$ be an infinite complete two-pattern string of scope 1 reducible by $\rho = [a, b, i, j]$, $i < j$. Then*

(a) *for $k \in 1..i$, $C_1(k) = k+1$;*

(b) *for $k \in i+1..j$, $2k-i \leq C_1(k) \leq 3k-(2i+1)$;*

(c) *for $k \geq j+1$, $k+(j-i) \leq C_1(k) \leq 3k-(2i+2)$.* □

See also [J01].

If $\boldsymbol{u}$ is a nonempty substring of a string $\boldsymbol{x}$ such that both $a\boldsymbol{u}$ and $b\boldsymbol{u}$ are prolific in $\boldsymbol{x}$, we say that $\boldsymbol{u}$ is ***special*** in $\boldsymbol{x}$. It is then easy to prove

Lemma 8 *Suppose that* $\boldsymbol{x}$ *is an infinite complete twopattern string of scope 1 reducible by* $\rho = [a, b, i, j]$, $i < j$. *Then* a^k *is special in* $\boldsymbol{x}$ *if and only if* $k = i$ *and* $j > i{+}1$. $\square$

Lemma 9 *Let* $\boldsymbol{u}$ *be a special substring of length* k *of an infinite complete two-pattern string* $\boldsymbol{x}$ *of scope 1. Then*

$$C_1(k{+}2) \geq C_1(k{+}1){+}2. \quad \square$$

To illustrate Lemma 9, consider again the example (9) with three prolific substrings (10): observe that in this case $\boldsymbol{u} = ababa$ of length 5 is special, so that $C_1(7) \geq C_1(6){+}2$.

Theorem 5 extends naturally to special strings, where now the exclusion of exceptional and left-extendible strings is no longer of interest, since neither of these can be special:

Theorem 10 *Suppose that an infinite complete two-pattern string* $\boldsymbol{x}$ *is reducible by* $\rho = [a, b, i, j]$, $i < j$. *Let* $\boldsymbol{u} = a^i b \boldsymbol{v} a^i$ *be a substring of* $\boldsymbol{x}$ *in standard form. Then*

$$\boldsymbol{u} \text{ special in } \boldsymbol{x} \Longleftrightarrow \rho(\boldsymbol{v}) \text{ special in } \rho(\boldsymbol{x}). \quad \square$$

For example, the special substring $\boldsymbol{u} = ababa$ in (9) reduces to the special substring a in $\rho(\boldsymbol{x})$.

We shall say that a reduction $\rho = [a, b, i, j]$ is ***Sturmian*** if $j = i{+}1$, otherwise ***non-Sturmian***.

Theorem 11 *Let* $\boldsymbol{x}$ *be an infinite complete itwo-pattern string of scope 1 reducible by* $\rho = [a, b, i, j]$, $i < j$. *The number of special substrings* $\boldsymbol{u}$ *in* $\boldsymbol{x}$ *is exactly equal to the number of non-Sturmian reductions of* $\boldsymbol{x}$.

Proof Suppose that $\boldsymbol{u}$ is special in $\boldsymbol{x}$, so that both $a\boldsymbol{u}$ and $b\boldsymbol{u}$ are prolific in $\boldsymbol{x}$. If $\boldsymbol{u} = a^i$, then by Lemma 8, ρ is non-Sturmian.

Suppose then that $\boldsymbol{u} \neq a^i$. By Lemma 3(b) $\boldsymbol{u}$ is also prolific, and moreover must have prefix $a^i b$. Furthermore by Lemma 3(a) $\boldsymbol{u}$ has suffix ba^i, and so is in standard form. It is easily verified that if $\boldsymbol{u} = a^i b a^i$, then $a\boldsymbol{u}$ and $b\boldsymbol{u}$ cannot both be prolific, and so we may suppose that $\boldsymbol{u} = a^i b \boldsymbol{v} b a^i$ for some nonempty $\boldsymbol{v}$.

Since $a\boldsymbol{u}$ is prolific, it follows from Lemma 3(c) that $\boldsymbol{u_a}a^i = a^iba^{j-i}\boldsymbol{u}$ is prolific; since $b\boldsymbol{u}$ is prolific, so also is $\boldsymbol{u_b}a^i = a^ib\boldsymbol{u}$. Because both $\boldsymbol{u_a}a^i$ and $\boldsymbol{u_b}a^i$ are in standard form, it follows from Theorem 5 that

$$\begin{aligned}\rho(\boldsymbol{u_a}) &= ab\rho(\boldsymbol{v}b),\\ \rho(\boldsymbol{u_b}) &= aa\rho(\boldsymbol{v}b)\end{aligned}$$

are both prolific in $\boldsymbol{x'} = \rho(\boldsymbol{x})$. Hence we have identified a nonempty string $\boldsymbol{u'} = \rho(\boldsymbol{v}b)$ such that both $a\boldsymbol{u'}$ and $b\boldsymbol{u'}$ are prolific in $\boldsymbol{x'}$, with $|\boldsymbol{u'}| < |\boldsymbol{u}|$. Thus $\boldsymbol{u'}$ is special in $\boldsymbol{x'}$. Either $\boldsymbol{u'} = a^{i'}$, where $\rho' = [a, b, i', j']$ is the reduction for $\boldsymbol{x'}$, or else the transformation can be repeated. Since $\boldsymbol{u}$ is of finite length, we must ultimately transform into $\boldsymbol{u'} = a^{i'}$, a special substring of an infinite complete two-pattern string of scope 1, say $\boldsymbol{x'}$, reducible by $\rho' = [a, b, i', j']$.

But by Lemma 8, $a^{i'}$ is special if and only if $j' > i'+1$; in other words, if and only if ρ' is non-Sturmian. Thus, corresponding to every special substring of $\boldsymbol{x}$, there exists a non-Sturmian reduction of $\boldsymbol{x}$.

Conversely, if $a^{i'}$ is special in any string, it must by Theorem 10 map into a special string $\boldsymbol{u}$ in $\boldsymbol{x}$. Thus, corresponding to every non-Sturmian reduction of $\boldsymbol{x}$, there exists a special substring of $\boldsymbol{x}$. □

When there exist only Sturmian reductions of $\boldsymbol{x}$, Theorem 11 tells us that there exist no special strings, and hence provides an alternate proof of the fact that for Sturmian strings, $C(k) = k+1$ for all k. But in view of Lemma 9, Theorem 11 has a much more significant consequence:

Theorem 12 *If an infinite sequence of reductions of an infinite complete two-pattern string $\boldsymbol{x}$ of scope 1 contains an infinite number of non-Sturmian reductions, then*

$$\lim_{k\to\infty}(C_1(k)-k) = \infty.$$

Proof By Theorem 11, corresponding to each non-Sturmian reduction $r = 1, 2, \ldots,$ there exists a substring of length k_r such that, by Lemma 9,

$$C_1(k_r) \geq C_1(k_r-1)+2.$$

Since every right extension of a distinct string is distinct, it follows that for sufficiently large k, $C_1(k)-k$ is unbounded. □

Of course this result holds also for scope $\lambda > 1$: the complexity of two-pattern strings can be arbitrarily large.

A more precise result is available in the case that the sequence of reductions of $\boldsymbol{x}$ contains only a finite number r of non-Sturmian reductions. Recall that by Lemma 3(a), every prolific substring $\boldsymbol{u} \neq a^k$, $k \in 1..j-1$, has suffix ba^i, itself a prolific substring of $\boldsymbol{x}$. Thus a new prolific substring of length $k+1$ can be formed only using an existing prolific substring of length k. As we have seen, it is the special substrings that provide the means of creating two distinct prolific substrings of length $k+1$ out of a single prolific substring of length k.

Suppose that $\boldsymbol{x}$ is reducible by $\rho = [a, b, i, j]$ to $\boldsymbol{y} = \rho(\boldsymbol{x})$, itself in turn reducible by a non-Sturmian reduction $\rho' = [a, b, i', j']$. Then in $\boldsymbol{x}$ there exists the special substring

$$\boldsymbol{u} = a^i b(a^i b)^{i'} a^i$$

of length $(i'+1)(i+1)+i$, giving rise to two prolific substrings

$$a\boldsymbol{u} = a^{i+1} b(a^i b)^{i'} a^i \quad \& \quad b\boldsymbol{u} = b(a^i b)^{i'+1} a^i$$

of length $(i'+2)(i+1)$. (Note that for $j' = i'+1$, $b\boldsymbol{u}$ is *not* prolific, and so $\boldsymbol{u}$ is not special.)

Considering first $a\boldsymbol{u}$, observe that the sequence

$$a^{i+1} b(a^i b)^{i'} a^i, \ldots, (a^i b)^{i'} a^j b(a^i b)^{i'} a^i$$

is entirely prolific (and indeed may perhaps be extended). Thus corresponding to $a\boldsymbol{u}$, there exists a sequence of at least $(i'-1)(i+1)+(j+1)$ prolific substrings in $\boldsymbol{x}$ of lengths $k \in (i'+2)(i+1)..(2i'+1)(i+1)+j$.

Considering $b\boldsymbol{u}$, we find that the sequence of substrings

$$b(a^i b)^{i'+1} a^i, \ldots, (a^i b)^{j'} a^i$$

is entirely prolific, while $\mu(a^i b)^{j'} a^i$ is not prolific for any $\mu \in \{a, b\}$, since

$$a^{i+1} b(a^i b)^{j'-1} a^i b \quad \& \quad b(a^i b)^{j'} a^{i+1}$$

are the only possibile extensions. Thus corresponding to $b\boldsymbol{u}$, there exists a sequence of exactly $(j'-i'-1)(i+1)$ prolific substrings in $\boldsymbol{x}$ of lengths $k \in (i'+2)(i+1)..j'(i+1)+i$.

Putting these two cases together, we have

Lemma 13 *Let $\boldsymbol{x}$ be an infinite complete two-pattern string of scope 1 reducible by $\rho = [a, b, i, j]$, $i < j$. Let $\rho(\boldsymbol{x})$ be reducible by $\rho' = [a, b, i', j']$, $j' > i'+1$. Then for $\boldsymbol{x}$,*

$$C_1(k+1) \geq C_1(k)+2$$

for every

$$k \in (i'+2)(i+1)..\min\big\{j'(i+1)+i, (2i'+1)(i+1)+j\big\}.$$

The minimum range of values of k is $i+1 \geq 2$, attained for $j' = i'+2$. □

Since the expansion of any collection of distinct substrings yields another distinct collection, we have

Theorem 14 *Let $\boldsymbol{x}$ be an infinite complete two-pattern string of scope 1 with a sequence of reductions containing exactly r non-Sturmian reductions. Then for sufficiently large k,*

$$C_1(k)-k \geq 4r+1.$$

Proof For each of the r reductions, there must by Lemma 13 be at least two consecutive values, say $k' \in k+1..k+2$, for which $C_1(k') \geq C_1(k'-1)+2$, so that $C_1(k+2) \geq C_1(k)+4$. The result follows. □

4 Conclusion & Open Problems

The most striking result of this paper is that the rather slight generalization of the Sturmian strings to infinite complete two-pattern strings of scope $\lambda = 1$ gives rise to strings whose substring complexity $C_1(k)$ can become arbitrarily large for arbitrarily long substrings of length k. Since for $\lambda = 1$ the only possible patterns are a and b, this means that the result holds quite independent of the elaborate definition of suitable patterns given in Section 1 for the general case $\lambda > 1$. In the case $\lambda = 1$, the only way to differ from the Sturmian case is to have non-Sturmian reductions (where $|j-i| > 1$). It follows that the Sturmian strings are, indeed, optimal with respect to minimality of substring complexity.

We believe that more precise complexity results can be formulated for scope 1 than we have been able to achieve in this paper. Also, it would be of interest to investigate the complexity of two-patterns strings in the general case.

References

[FKS00] Frantisek Franek, Ayşe Karaman & W. F. Smyth, **Repetitions in Sturmian strings**, *TCS 249–2* (2000) 289–303.

[FLS03] Frantisek Franek, Weilin Lu & W. F. Smyth, **Two-pattern strings I — a recognition algorithm**, *J. Discrete Algs. 1–5/6* (2003) 445–460.

[FLS04] Frantisek Franek, Weilin Lu & W. F. Smyth, **Two-pattern strings II — Computing Repetitions & Near-Repetitions**, submitted for publication.

[GO81] Leo J. Guibas & Andrew M. Odlyzko, **Periods in strings**, *J. Combinatorial Th. Series A 30* (1981) 19–42.

[J01] Jiandong Jiang, *Frequency of Occurrence of Two-Pattern Strings*, M.Sc. thesis, McMaster University (2001).

[L02] Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press (2002) 504 pp.

[S03] Bill Smyth, *Computing Patterns in Strings*, Pearson Addison-Wesley (2003) 423 pp.

Acknowledgements

The authors acknowledge the support provided by the Natural Sciences & Engineering Research Council of Canada.