A Combinatorial View on String Attractors[☆]

Sabrina Mantaci[a], Antonio Restivo[a], Giuseppe Romana[a], Giovanna Rosone[b], Marinella Sciortino[a]

[a] *University of Palermo, Italy*
[b] *University of Pisa, Italy*

Abstract

The notion of *string attractor* has recently been introduced in [Prezza, 2017] and studied in [Kempa and Prezza, 2018] to provide a unifying framework for known dictionary-based compressors. A string attractor for a word $w = w_1 w_2 \cdots w_n$ is a subset Γ of the positions $\{1, \ldots, n\}$, such that all distinct factors of w have an occurrence crossing at least one of the elements of Γ.

In this paper we explore the notion of string attractor by focusing on its combinatorial properties. In particular, we show how the size of the smallest string attractor of a word varies when combinatorial operations are applied and we deduce that such a measure is not monotone. Moreover, we introduce a circular variant of the notion of string attractor to provide a characterization of the conjugacy classes of standard Sturmian words.

Keywords: String attractor, Burrows-Wheeler transform, Lempel-Ziv encoding, standard Sturmian word, Thue-Morse word, de Bruijn word

Email addresses: `sabrina.mantaci@unipa.it` (Sabrina Mantaci), `antonio.restivo@unipa.it` (Antonio Restivo), `giuseppe.romana01@unipa.it` (Giuseppe Romana), `giovanna.rosone@unipi.it` (Giovanna Rosone), `marinella.sciortino@unipa.it` (Marinella Sciortino)

Postprint version; to appear on Theoretical Computer Science

1. Introduction

This paper focuses on the notion of *string attractor* that has been recently introduced and studied in [32, 16] to find a common principle underlying the main techniques constituting the fields of dictionary-based compression. It is defined as a subset of the text's positions such that all distinct factors have an occurrence crossing at least one of the string attractor's elements. On one hand the problem of finding the smallest string attractor of a word has been proved to be NP-complete, on the other hand most well-known compression schemes, such as straight-line programs, Run-Length Burrows-Wheeler transform, macro schemes, collage systems, and the compact directed acyclic word graphs, can be interpreted as algorithms approximating the smallest string attractor for a given word [16]. In particular, the size of the string attractors induced by the compression schemes can be bounded by the repetitiveness measures associated to such compressors. This fact can allow to discover asymptotic relations between the output sizes of different compressors (cf. [17]), with several applications in designing new data structures and techniques, especially for indexing compressed massive and highly-repetitive data (see for instance [6, 13] and references therein).

In this paper we use some results related to the compressors based on the Burrows-Wheeler Transform and the dictionary-based compressors.

The Burrows-Wheeler Transform (BWT) is a reversible transformation that was introduced in 1994 in the field of Data Compression and it is also largely used for self-indexing data structures. It has several combinatorial properties that make it a versatile tool in several contexts and applications [33, 34, 24, 30, 35, 29, 15].

Dictionary-based compressors are mainly based on a technique originated in two theoretical papers of Ziv and Lempel [37, 38]. Such compressors, that are able to combine compression power and compression/decompression speed, are based on a paper in which combinatorial properties of word factorization are explored [22]. The relationship between LZ77 and BWT has been investigated from the algorithmic point of view in [31].

The main goal of this paper is to explore the combinatorial properties of string attractors. More in detail, we are interested in how the size of the smallest string attractor of a word varies when combinatorial operations are applied. We also show that one of the consequences of these combinatorial properties is that the complexity measure defined by the size of the string attractors is not monotone. This answers to an open problem posed in [20].

Furthermore, we are interested to consider the problem of computing string attractors for infinite families of words that are well known in the field of Combinatorics on Words: standard Sturmian words, Thue-Morse words and de Bruijn words. In particular, we prove that the size of the smallest string attractor for standard Sturmian words is 2 and it contains two consecutive positions. For the de Bruijn words the size of the smallest string attractor grows asymptotically as $\frac{n}{\log n}$, where n is the length of the word. In [21] it has been proved that Thue-Morse words have a smallest string attractor of size 4.

Finally, a circular variant of the notion of string attractor is here introduced to characterize the conjugacy classes of standard Sturmian words. This notion may have an interest independent of this result.

A preliminary version of the results here presented can be found in [27].

2. Preliminaries

Let $\Sigma = \{a_1, a_2, \ldots, a_\sigma\}$ be a finite ordered alphabet with $a_1 < a_2 < \ldots < a_\sigma$, where $<$ denotes the standard lexicographic order. We denote by Σ^* the set of words over Σ. Given a finite word $w = w_1 w_2 \cdots w_n \in \Sigma^*$ with each $w_i \in \Sigma$, the length of w, denoted $|w|$, is equal to n.

Given a finite word $w = w_1 w_2 \cdots w_n$ with each $w_i \in \Sigma$, a *factor* of a word w is written as $w[i, j] = w_i \cdots w_j$ with $1 \leq i \leq j \leq n$. A factor of type $w[1, j]$ is called a *prefix*, while a factor of type $w[i, n]$ is called a *suffix*. We also denote by $w[i]$ the i-th letter in w for any $1 \leq i \leq n$. Let us denote by $f_w(k)$ the number of distinct factors of w having length k. The function f_w is called *factor complexity* of w. An *(equal-letter) run* in a word w is a maximal factor a^k, with $k > 0$ and $a \in \Sigma$. Therefore, a word w can be uniquely written as $w = u_1 u_2 \cdots u_r$, where u_i are equal-letter runs of w and r is called *number of runs* of w.

We denote by $\overleftarrow{w}$ the reversal of w, given by $\overleftarrow{w} = w_n \cdots w_2 w_1$. If w is a word that has the property of reading the same in either direction, i.e. if $w = \overleftarrow{w}$, then w is called a *palindrome*.

We say that two words $x, y \in \Sigma^*$ are *conjugate*, if $x = uv$ and $y = vu$, where $u, v \in \Sigma^*$. Conjugacy between words is an equivalence relation over Σ^*. We say that u has a *circular occurrence* in w if u is a factor of a conjugate of w. In this case we say that u is a *circular factor* of w. More formally, an occurrence of a circular factor v of w is identified by a pair (i, j), with

$i, j \in \{1, 2, \ldots, n\}$, such that

$$v = \begin{cases} w[i, j] & \text{if } i \leq j \\ w[i, n] \cdot w[1, j] & \text{if } i > j. \end{cases}$$

For instance aa is a circular factor of $abbbba$, but it is not a factor. Let us denote by $c_w(k)$ the number of distinct circular factors of w having length k. The function c_w is called *circular factor complexity* of w. It is easy to see that $f_w(k) \leq c_w(k)$, for each $k \geq 1$.

Given a finite word w, w^k denotes the word obtained by concatenating k copies of w. A nonempty word $w \in \Sigma^+$ is *primitive* if $w = u^h$ implies $w = u$ and $h = 1$. Given a word x, a positive integer $p \leq |x|$ is a *period* of x if $x[i] = x[j]$ when $i = j \mod p$.

The *Burrows-Wheeler Transform* (BWT) is a word transformation and it was introduced in the field of Data Compression [8]. More formally, given a word $w \in \Sigma^*$, the output of BWT is the pair $(\texttt{bwt}(w), I)$, where $\texttt{bwt}(w)$ is the permutation of the letters in the input word w obtained by considering the matrix M containing the lexicographically sorted list of the cyclic rotations of w, and by concatenating the letters of the last column L of M; I is the position where the original word w appears in M. An important property that assures the reversibility of BWT is that for each letter c, the i-th occurrence of c in the last column L of the matrix M corresponds to the i-th occurrence of c in its first column F. Such a correspondence is called *LF-mapping*. In fact, given the output of BWT, the original word w can be recovered as follows: $w[n-k] = L[LF^k[I]]$ for $0 \leq k \leq n-1$. Note that if an end-of-string symbol $\$ \notin \Sigma$ (and smaller than any symbol in Σ) is appended to the word w, lexicographically sorting the cyclic rotations of $w\$$ can be reduced to sorting its suffixes [14, 3, 4, 28]. Therefore, adding the \$ symbol at the end of w changes the output of BWT compared with $BWT(w)$. In fact, as shown in Fig. 1, both the index I and the words $\texttt{bwt}(w)$ and $\texttt{bwt}(w\$)$ may be quite different.

The *LZ-parsing* of a word w is its factorization $s = p_1 \cdots p_z$ built left to right in a greedy way by the following rule: each new factor (also called an LZ-phrase) p_i is either the leftmost occurrence of a letter in w or the longest prefix of $p_i \cdots p_z$ which occurs, as a factor, in $p_1 \cdots p_{i-1}$.

	F										L
	↓										↓
1	a	a	a	a	a	b	a	a	a	a	b
2	a	a	a	a	b	a	a	a	a	a	b
3	a	a	a	a	b	a	a	a	a	b	a
4	a	a	a	b	a	a	a	a	a	b	a
5	a	a	a	b	a	a	a	a	b	a	a
6	a	a	b	a	a	a	a	a	b	a	a
7	a	a	b	a	a	a	a	b	a	a	a
8	a	b	a	a	a	a	a	b	a	a	a
9	a	b	a	a	a	a	b	a	a	a	a
10	b	a	a	a	a	a	b	a	a	a	a
11	b	a	a	a	a	b	a	a	a	a	a

(a)

	F										L
	↓										↓
1	\$	a	a	a	b	a	a	a	a	a	b	a
2	a	\$	a	a	a	b	a	a	a	a	a	b
3	a	a	a	a	a	b	a	\$	a	a	a	b
4	a	a	a	a	b	a	\$	a	a	a	b	a
5	a	a	a	b	a	\$	a	a	a	b	a	a
6	a	a	a	b	a	a	a	a	a	b	a	\$
7	a	a	b	a	\$	a	a	a	b	a	a	a
8	a	a	b	a	a	a	a	a	b	a	\$	a
9	a	b	a	\$	a	a	a	b	a	a	a	a
10	a	b	a	a	a	a	a	b	a	\$	a	a
11	b	a	\$	a	a	a	b	a	a	a	a	a
12	b	a	a	a	a	a	b	a	\$	a	a	a

(b)

Figure 1: (a) The matrix lexicographically sorted cyclic rotations of the word $w = aaabaaaaaba$. The last column of the matrix is $\texttt{bwt}(w) = bbaaaaaaaaa$ and $I = 4$. (b) The matrix of lexicographically sorted cyclic rotations of the word $w\$$. Then, $\texttt{bwt}(w\$) = abbaa\$aaaaaa$ and $I = 6$.

3. String Attractor of a word

In this section we describe the notion of string attractor that is a combinatorial object introduced in [32, 16] to obtain a unifying framework for dictionary compressors.

Definition 1. A *string attractor* of a word $w \in \Sigma^n$ is a set of γ positions $\Gamma = \{j_1, \ldots, j_\gamma\}$ such that every factor $w[i, j]$ has an occurrence $w[i', j'] = w[i, j]$ with $j_k \in [i', j']$, for some $j_k \in \Gamma$.

Simply put, a string attractor for a word w is a set of positions in w such that all distinct factors of w have an occurrence *crossing* at least one of the attractor's elements. Note that, trivially, any set that contains a string attractor for w, is a string attractor for w as well. Note also that a word can have different string attractors that are not included into each other. We are interested in finding a *smallest string attractor*, i.e. a string attractor with a minimum number of elements. We denote by $\gamma^*(w)$ the size of the smallest

string attractor for w. Note that all the factors made of a single letter should be covered, and therefore $\gamma^*(w) \geq |\Sigma|$.

Example 2. *Let $w = adcbaadcbadc$ be a word on the alphabet $\Sigma = \{a, b, c, d\}$. A string attractor for w is for instance $\Gamma = \{1, 4, 6, 8, 11\}$. Note that position 1 can be removed from Γ, since all the factors that cross position 1 have a different occurrence that crosses a different position in Γ. Therefore $\Gamma' = \{4, 6, 8, 11\}$ is also a string attractor for w with a smaller number of elements. The positions of Γ' are underlined in*

$$w = adc\underline{b}a\underline{a}d\underline{c}ba\underline{d}c.$$

Γ' is also a smallest string attractor since $|\Gamma'| = |\Sigma|$. Then $\gamma^*(w) = 4$. Remark that the sets $\{3, 4, 5, 11\}$ and $\{3, 4, 6, 7, 11\}$ are also string attractors for w. It is easy to verify that the set $\Delta = \{1, 2, 3, 4\}$ is not a string attractor since, for instance, the factor aa does not intersect any position in Δ.

In [16] the authors show that many of the most well-known compression schemes reducing the texts size by exploiting its repetitiveness can induce string attractors whose sizes are bounded by the repetitiveness measures associated to such compressors. In particular, straight-line programs, Run-Length Burrows-Wheeler transform, macro schemes, collage systems, and the compact directed acyclic word graph are considered. Here we report some results related to the Burrows-Wheeler transform and Lempel-Ziv 77 (that is a particular macro-scheme) that provide upper bounds on the size of the smallest string attractor for a given word. Such bounds will be used in next sections to compute the string attractors for known infinite families of words.

Let v be a word of length n and let $BWT(v) = (\texttt{bwt}(v), I)$. Let us denote by $\Gamma_{\texttt{bwt}}(v)$ the set

$$\{n - k \mid LF^k[I] = 1 \text{ or } L[LF^k[I] - 1] \neq L[LF^k[I]\},$$

i.e. $\Gamma_{\texttt{bwt}}$ is the set of positions of the symbols in v that correspond to the first occurrence of a symbol in each equal-letter run in $\texttt{bwt}(v)$ (alternatively, we can consider the set $\Gamma'_{\texttt{bwt}}$ defined as the set of positions at the end of equal-letter runs in $\texttt{bwt}(v)$).

The following theorem, proved in [16], states a relation between a string attractor of a word and the runs of its `bwt` when a \$-symbol is appended.

Theorem 3 ([16]). *Let Σ be a finite alphabet and $\$ \notin \Sigma$ is a symbol smaller than any symbol in Σ. Let $w \in \Sigma^*$ and r be the number of equal-letter runs in the $\mathtt{bwt}(w\$)$. Then, $\Gamma_{\mathtt{bwt}}(w\$)$ is a string attractor for $w\$$ and $\gamma^*(w\$) \leq r$.*

Example 4. *Let us consider the word $w = aaabaaaaaba$. The lexicographically sorted cyclic rotations of*

		1	2	3	4	**5**	6	**7**	8	9	**10**	**11**	12
$w\$$	=	a	a	a	b	***a***	a	***a***	a	a	***b***	***a***	***$***

are shown in Figure 1(b). By applying Theorem 3 we can construct the string attractor $\{5, 7, 10, 11\}$ for w obtained by considering the position in w (in bold) of the symbols appearing at the beginning of each equal-letters run and by removing the position 12 from $\Gamma_{\mathtt{bwt}}(w\$)$.

Remark 5. *In general, for a given word w, the string attractor constructed by using Theorem 3 is not necessarily the smallest one. For instance, it is easy to see that the string attractor defined in Example 4 for $w = aaabaaaaaba$ is not the smallest one, because the set $\{4, 9\}$ is a string attractor too. Note also that if the \$-symbol is not appended to the word w, the positions that correspond to the symbols at the beginning of equal-letter runs in $\mathtt{bwt}(w)$ (or alternatively the symbols at the end of equal-letter runs in $\mathtt{bwt}(w)$) is not in general a string attractor. In fact, the set $\{4, 5\}$ (corresponding to the positions in w of the first b and the first a in the runs of $bwt(w)$) is not a string attractor for w since the factor $aaaab$ has no occurrence crossing any position in such a set.*

When the \$-symbol is not used, a result analogous to Theorem 3 can be obtained if the occurrences of circular factors are considered.

Given a word w of length n and a set $\Gamma \subseteq [1, n]$, we say that the occurrence of a circular factor v of w, specified by the pair (i, j), *crosses* a position $p \in \Gamma$ if

$$\begin{cases} p \in [i, j] & \text{if } i \leq j \\ p \in [i, n] \cup [1, j] & \text{if } i > j. \end{cases}$$

By using the previous definition and adapting the proof of Theorem 3 to the context of circular factors, we can derive the following:

Theorem 6. *Let Σ be a finite alphabet, $w \in \Sigma^*$ and r be the number of equal-letter runs in the $\mathtt{bwt}(w)$. Then, each circular factor of w has a occurrence that crosses a position in $\Gamma_{\mathtt{bwt}}(w)$.*

PROOF. To prove that each circular factor u of w has an occurrence (i, j) crossing at least a position in $\Gamma_{\texttt{bwt}}$, consider the index $J = LF^{n-j}[I]$, i.e. the index of the cyclic rotation such that $w[j]$ corresponds to $L[J]$ in M. Moreover, let ℓ be the length of u, and let $[l_0, r_0], [l_1, r_1], \ldots, [l_{\ell-1}, r_{\ell-1}]$ be the sequence of equal-letter runs visited in the column L while applying the *LF-mapping* from $w[j]$ to $w[i]$, i.e. $L[l_t, r_t]$ contains $L[LF^t[J]]$ for any $t \in [0, \ell - 1]$. Consider the value $\Delta = \min\{LF^t[J] - l_t | t \in [0, \ell - 1]\}$. Recall that $\Gamma_{\texttt{bwt}} = \{p_s | w[p_s]$ corresponds to $L[l_s]$ for any $s \in [0, r - 1]\}$. Hence, if $\Delta = LF^m[J] - l_m = 0$ for some $m \in [0, \ell - 1]$, then $LF^m[J] = l_m$ and the occurrence (i, j) of u has a symbol which crosses a position in $\Gamma_{\texttt{bwt}}$. Otherwise, assume $\Delta = LF^m[J] - l_m > 0$. It is easy to see that if two positions p_1, p_2 belong to the same equal-letter run in L then $L[p_2] = L[p_1] + (p_2 - p_1)$. Thus, if we pick $J' = J - \Delta$, at any step we have $L[LF^t[J']] = L[LF^t[J]]$ for any $t \in [0, \ell - 1]$, which means that there exists another circular occurrence (i', j') of u, where $J' = LF^{n-j'}[I]$. Since $w[LF^m[J']]$ corresponds to $L[l_m]$, the circular occurrence (i', j') contains a position from $\Gamma_{\texttt{bwt}}$. □

Example 7. *Let us consider the word $w = aaabaaaaaba$. The lexicographically sorted cyclic rotations of*

		1	2	3	**4**	**5**	6	7	8	9	10	11
w	=	a	a	a	***b***	***a***	a	a	a	a	b	a

are shown in Figure 1(a). By applying Theorem 6 we can construct the set $\{4, 5\}$ for w obtained by considering the position in w of the symbols (in bold) appearing at the beginning of each equal-letters run. As shown in Remark 5, the factor $w[6, 10] = aaaab$ has not any occurrence crossing the set $\{4, 5\}$, but its circular occurrence $(11, 4)$ crosses position 4.

The following result, proved in [16], states the relationship between a string attractor of a word w and the number of phrases in the LZ-parsing of w. In particular, a string attractor can be constructed by considering the set of positions at the end of each phrase.

Theorem 8 ([16]). *Given a word w, there exists a string attractor of w of size equal to the number of phrases of its LZ-parsing.*

4. Combinatorial properties of string attractors

In this section we explore some combinatorial properties of the string attractors and in particular how the sizes of smallest string attractors are affected by the application of different operations on strings.

The following two propositions, proved in [32], are useful in order to derive a lower bound on the value of γ^*.

Proposition 9 ([32]). *Let Γ be a string attractor for the word w. Then, $f_w(k) \leq |\Gamma|k$, for every $1 \leq k \leq |w|$.*

Proposition 10 ([32]). *Let $w \in \Sigma^*$ and let ℓ be the length of its longest repeated factor. Then it holds $\gamma^*(w) \geq \frac{|w|-\ell}{r+1}$.*

The next proposition states the relation between the string attractor of a word and a string attractor of its reverse.

Proposition 11. *Let w be a word and let $\overleftarrow{w}$ denote its reverse. Then, $\gamma^*(w) = \gamma^*(\overleftarrow{w})$.*

Proof. Let $\Gamma = \{p_1, p_2, \ldots, p_\gamma \mid 1 \leq p_i \leq |w|, 1 \leq i \leq \gamma\}$ be a string attractor for w and consider the corresponding positions $\overleftarrow{\Gamma} = \{n - p_\gamma + 1, n - p_{\gamma-1} + 1, \ldots, n - p_1 + 1\}$ in $\overleftarrow{w}$. Let v be any factor of $\overleftarrow{w}$. Then its reverse $\overleftarrow{v}$ is a factor of w. Since Γ is a string attractor for w, then there exists a position $p_i \in \Gamma$ that intercepts an occurrence of $\overleftarrow{v}$. Therefore an occurrence of v is intercepted by $n - p_i + 1 \in \overleftarrow{\Gamma}$, i.e. $\overleftarrow{\Gamma}$ is a string attractor for $\overleftarrow{w}$. In particular if Γ is a smallest string attractor for w, then $\overleftarrow{\Gamma}$ is a smallest string attractor for $\overleftarrow{w}$, since otherwise we could find a smaller string attractor for w, in contradiction with the hypothesis of minimality. □

When a word w is obtained as a concatenation of two factors u and v, an upper bound for $\gamma^*(w)$ can be expressed in terms of $\gamma^*(u)$ and $\gamma^*(v)$, as stated in the following theorem.

Proposition 12. *Let u and v two words, then $\gamma^*(uv) \leq \gamma^*(u) + \gamma^*(v) + 1$.*

Proof. Let $\Gamma^*(u)$ and $\Gamma^*(v)$ be smallest string attractors for u and v, respectively. Then $\Gamma^*(u) \cup \{p + |u| \mid p \in \Gamma^*(v)\}$ covers all the factors of u and v but might not cover some of new factors that appear across the concatenation point of u and v. In this case it is sufficient to add the last position of the prefix u (or the first position of the suffix v) to have a string attractor for uv. □

Example 13. *The bound defined in the previous proposition is tight. In fact, let* $u = \underline{b}aa\underline{a}ba$ *and* $v = c\underline{d}cc\underline{c}d$ *be two words in which the positions of the respective smallest string attractors are underlined. If we consider the concatenation* $uv = \underline{b}aa\underline{a}ba\underline{cd}cc\underline{c}d$*, the set underlined positions represent one of the smallest string attractors for* uv*, as one can verify, having cardinality 5.*

Although γ^* is sensitive to the concatenation operation, the following proposition shows that it is not a monotone measure, in the sense that there exist words $w = uv$ such that $\gamma^*(u) > \gamma^*(w)$. This answers to a problem posed in [20].

Proposition 14. *The measure* γ^* *is not monotone.*

Proof. We show the statement by showing an example where monotonicity does not hold. For each $n > 0$, let us consider $w = a\underline{b}b\underline{b}a^n\underline{a}b$. In this case $\gamma^*(w) = 3$ since by exhaustive search $\{2, 4, n+5\}$ is a smallest string attractor for w. On the other hand it is easy to verify that $\{4, n+5\}$ is a string attractor for $wb = abb\underline{b}a^n\underline{a}bb$, then $\gamma^*(wb) = 2$. □

The following proposition gives an upper bound for γ^*, when a power of a given word is considered.

Proposition 15. *Let* w *a word over the alphabet* Σ*. Then* $\gamma^*(w^n) \leq \gamma^*(w)+1$*. Moreover,* $\gamma^*(w^n) = \gamma^*(w^2)$ *for any* $n \geq 2$*.*

Proof. Let $\Gamma^*(w)$ denote a smallest string attractor for w. The positions of $\Gamma^*(w)$ cover all the factors that are contained in each occurrence of w in w^n, then no further position is needed to cover these factors. The only factors of w^n that might not be covered by $\Gamma^*(w)$ are the ones across the end of the first occurrence of w and the beginning of the following occurrence. Therefore it is sufficient to add the last position of the first (or any) occurrence of w in order to cover all these factors.

To prove that $\gamma^*(w^n) = \gamma^*(w^2)$ for any $n \geq 2$, we show that for any smallest string attractor for $\Gamma^*(w^n)$ we can deduce a string attractor of at most the same size for w^2 and the other way around. Given a smallest string attractor $\Gamma^*(w^n)$ with $n > 2$, we can create the set $\Delta = \{1 + (p-1) \bmod \ell \mid p \in \Gamma^*(w^n)\}$, where $\ell = |w|$ (note that $|\Delta| \leq \gamma^*(w^n)$). Consider the set $\Gamma = \Delta \setminus \{p_1\} \cup \{p_1 + \ell\}$, where $p_1 = \min \Delta$, i.e. Γ contains the positions of

Δ with only its leftmost position moved to the second occurrence of w. We now show that Γ is a string attractor for w^2. All the factors u that do not have occurrences overlapping two consecutive w's are covered in w^n by some position $p' \in \Gamma^*(w^n)$, i.e. $p' \in [i,j]$ where $w^n[i,j] = u$. Since ℓ is a period of w^n, it is easy to see that $w[1+(i-1) \mod \ell, 1+(j-1) \mod \ell)] = u$, hence either u crosses $1+(p'-1) \mod \ell$ in the first occurrence of w or, if $1+(p'-1) \mod \ell = p_1$, u crosses $(p_1 + \ell)$. For all the factors $u = w[i,\ell] \cdot w[1,j]$ that overlap two consecutive w's, if $i \leq p_\gamma$ or $j \geq p_1$ (where $p_\gamma = \max \Delta$), then u crosses p_γ or $(p_1 + \ell)$ respectively, so let us assume $i > p_\gamma$ and $j < p_1$. By construction of Γ we can deduce that u occurs also in $w^n[i',j']$ with $(i' \mod \ell) \neq (i \mod \ell)$ and $(j' \mod \ell) \neq (j \mod \ell)$ which crosses a position in $\Gamma^*(w^n)$. Given the periodicity of w^n, it is easy to see that either u occurs in w or u occurs again overlapping two consecutive w's. In both cases u crosses one of the positions in Γ.

Let now $\Gamma^*(w^2)$ be a smallest string attractor and consider the set $\Gamma' = \Gamma^*(w^2)$ if all of its positions lie within the second occurrence of w, otherwise $\Gamma' = \{p + \ell \mid p \in \Gamma^*(w^2)\}$. Since w^n has period ℓ, it is easy to see that the positions in Γ' point to the same symbols of $\Gamma^*(w^2)$, possibly moved in the following occurrence of w. Therefore, all the factors of w^n that are also factors of w^2 are already covered. The remaining factors are of the type $u = w[i,\ell] \cdot w^k \cdot w[1,j]$, with $k \in [1, n-2]$, $1 \leq i,j \leq \ell$. In this case, there is also an occurrence of u starting at the first or at the second w. Therefore, since all the positions of Γ' lie within the second or possibly within the third occurrence of w, u has to cross a position in Γ'. □

Example 16. *The bound given by Proposition 15 is tight. In fact consider the word $u = abbaab$. It is easy to check that the only smallest string attractors for u are $\Gamma_1 = \{2,4\}$ and $\Gamma_2 = \{3,5\}$.*

In order to find a smallest string attractor for $u^2 = abbaababbaab$, we remark that neither Γ_1 nor Γ_2 (neither any string attractor obtained from them by moving some position from the first to the second occurrence of u) cover all the new factors that appear after the concatenation. A way to get a smallest string attractor for u^2 is to add to Γ_1 or Γ_2, the position corresponding either to the end of the first occurrence of u or the beginning of the second occurrence. For instance, $\Gamma = \{2,4,6\}$ is a smallest string attractor for u^2.

Example 17. *We show that $\gamma^*(u^n)$ can be equal to $\gamma^*(u)$ although different points for the string attractor may have to be chosen. For instance, let*

$u = \underline{aba}\underline{b}\underline{c}bc$ be a word whose smallest string attractor is $\{2,3,5\}$ (the underlined letters). Then $u^2 = ab\underline{a}bc\underline{bc}ababcbc$ has a string attractor $\{3,6,7\}$ of cardinality 3. Remark that $\{2,3,5\}$ is not a string attractor for u^2.

The following proposition rectifies what stated in [27], i.e. shows that $\gamma^*(u) - \gamma^*(u^n)$ could become arbitrarily large.

Proposition 18. *For each $t > 0$, there exists an alphabet Σ_t and a word $w_t \in \Sigma_t^*$, such that $\gamma^*(w_t) - \gamma^*(w_t^n) > t$, for each $n \geq 2$.*

Proof. Let us consider $m = t+3$. We can define the string $w_t = v_1v_2v_3v_4v_5$ over the alphabet $\Sigma_t = \{a,b,c,d\} \cup \{\$_1, \$_2, \ldots, \$_{2m-1}\}$ such that

$$\begin{aligned}
v_1 &= c^{m-2}d^{m-1}\underline{\$_1},\\
v_2 &= a^{m-1}\underline{a}b^{m-1}\underline{b}c^{m-1}\underline{c}d^{m-1}\underline{d},\\
v_3 &= \Pi_{k=2}^{m-1}\underline{\$_k}a^{m-k}b^{m-1}\underline{c}c^{k-1},\\
v_4 &= \Pi_{k=1}^{m-1}\underline{\$_{m-1+k}}b^{m-k}\underline{c}c^{m-2}d^k,\\
v_5 &= \underline{\$_{2m-1}}a^{m-1}b^{m-1}\underline{c},
\end{aligned}$$

where Π denotes the concatenation of a set of words. The positions in a smallest string attractor for w_t are underlined. We note that the factors $a^{m-1}b^{m-1}c$, $a^{m-j}b^{m-1}c^j$, with $2 \leq j \leq m-1$, and $b^{m-j}c^{m-1}d^j$, with $1 \leq j \leq m-1$, appear once in v_5, v_3 and v_4, respectively. So, $\gamma^*(w_t) = 4m+1$. If we consider w_t^2, all the above mentioned factors occur in v_5v_1 and they are crossed by the rightmost position of v_5. Moreover, every other factor that appears in v_3, v_4 or v_5 which does not contain any $\$_i$ symbol, has another occurrence that is crossed either by the rightmost position in v_5 or by one of the underlined positions in v_2. Therefore, in order to obtain a smallest string attractor for w_t^2 we can remove the rightmost positions from each block of v_3 and v_4. Then, $\gamma^*(w_t^2) = 4m+1-(2m-3) = 2m+4$. This means that, by Proposition 15, $\gamma^*(w_t) - \gamma^*(w_t^n) = \gamma^*(w_t) - \gamma^*(w_t^2) = 4m+1-(2m+4) = 2m-3 = 2t+3 > t$. □

A consequence of previous proposition is that the difference between the string attractors of two conjugates can be arbitrarily large.

Corollary 19. *For each $t > 0$, there exists an alphabet Σ_t and a word $w_t = uv \in \Sigma_t^*$, such that $\gamma^*(uv) - \gamma^*(vu) > t$.*

Proof. The thesis follows by considering the word $w_t = v_1v_2v_3v_4v_5$ defined in the proof of Proposition 18 and its conjugate $w_t' = v_2v_3v_4v_5v_1$. □

5. String Attractors in Standard Sturmian words

In this section we explore a relationship between some combinatorial properties of strings and the structure of the correspondent smallest string attractor. In particular, we consider *standard Sturmian words* [23]. They represent a very well known family of binary words that are the basic bricks used for the construction of infinite Sturmian words, in the sense that every characteristic Sturmian word is the limit of an infinite sequence of standard Sturmian words (cf. Chapter 2 of [23]). These words have several characterizations and appear as extreme case in a very great range of contexts [19, 25, 26, 36, 11, 9, 1, 10, 12]. More formally, standard Sturmian words can be defined in the following way which is a natural generalization of the definition of the Fibonacci word.

Let $q_0, q_1, \ldots q_n, \ldots$ be any sequence of natural integers such that $q_0 \geq 0$ and $q_i > 0$ $(i = 1, \ldots, n, \ldots)$, called *directive sequence*. The sequence $\{s_n\}_{n\geq 0}$ can be defined inductively as follows: $s_0 = b$, $s_1 = a$, $s_{i+1} = (s_i)^{q_{i-1}} s_{i-1}$, for $i \geq 1$. We denote by *Stand* the set of all words s_n, $n \geq 0$, constructed for any directive sequence of integers. Such words are called *standard Sturmian words*.

A characterization of standard Sturmian words is related to the Burrows Wheeler transform (BWT) since, for binary alphabets, the application of the BWT to standard Sturmian words produces a total clustering of all the instances of any character (cf. [30]), as reported in the following theorem.

Theorem 20 ([30]). *Let $w \in \{a, b\}^*$. Then w is a conjugate of a word in Stand if and only if $\mathtt{bwt}(w) = b^p a^q$ with $gcd(p, q) = 1$.*

Remark 21. *Note that, as already shown in Figure 1, if we append a \$-symbol to a conjugate v of a word in Stand, the number of equal-letter runs of $\mathtt{bwt}(v\$)$ can be greater than 3. Moreover, it may not be the same for different conjugates. For instance, consider the standard Sturmian word s = ababaababaababababa and its conjugates t = ababaababaababaab (that also belongs to Stand) and v = baababababaababaaba. One can verify that $\mathtt{bwt}(s\$)$ = abbbbbbb\$aaaaaaaaa has 4 runs, $\mathtt{bwt}(t\$)$ = bbbbabbbaa\$aaaaaaa has 6 equal-letter runs and $\mathtt{bwt}(v\$)$ = abbbbabbabaaaa\$aaa has 9 equal-letter runs. By using Theorem 3, we can deduce that, for each of the above mentioned strings, it is possible to construct string attractors $\Gamma_{\mathtt{bwt}}$ with different size, i.e. $|\Gamma_{\mathtt{bwt}}(s\$)| = 4$, $|\Gamma_{\mathtt{bwt}}(t\$)| = 6$, $|\Gamma_{\mathtt{bwt}}(v\$)| = 9$.*

5.1. Minimum size string attractors

In this subsection, we study the problem of finding a smallest string attractor for the infinite family of standard Sturmian words. In particular, the following Theorem 22 shows that, for each standard Sturmian word, it is possible to find a string attractor having cardinality 2, whose positions are strictly related with particular decompositions of such words depending on their periodicity. In particular, we recall that $Stand = \{a, b\} \cup PER\{ab, ba\}$ (cf. [25]), where PER is the set of all words v having two periods p and q such that $gcd(p, q) = 1$ and $|v| = p + q - 2$. Given a word $w \in Stand$, we denote by $\pi(w)$ its prefix of length $|w| - 2$, belonging to the set PER, uniquely defined by using previous equality. By using a property of words in PER (cf. [25]), $\pi(w) = QxyP = PyxQ$, where $x \neq y$ are characters and Q and P are uniquely determined palindromes. So, a standard Sturmian word $w = \pi(w)ba$ can be decomposed as $w = QxyPba = PyxQba$. We call PER-decompositions such factorizations of w.

Theorem 22. *For each $w \in Stand$ with $|w| \geq 2$, let η be the length of the longest palindromic proper prefix of $\pi(w)$, the set $\Gamma_1 = \{\eta + 1, \eta + 2\}$ or the set $\Gamma_2 = \{|w| - \eta - 3, |w| - \eta - 2\}$ is a smallest string attractor for w.*

Proof. Let us suppose that $w = \pi(w)ba$. By using PER-decompositions, a Standard sturmian word can be decomposed as $w = QxyPba = PyxQba$, $\pi(w) = QxyP = PyxQ$, where $x \neq y$ are characters and Q and P are uniquely determined palindromes. Let us suppose that $|Q| > |P|$. So, $\eta = |Q|$. Firstly we suppose the case $x = b$. This means that $w = QbaPba = PabQba$. From a result in [2] $aPabQb$ and $bQbaPa$ are the smallest and the greatest conjugates in the lexicographic order, respectively. Let us consider the set $\Gamma_2 = \{|w| - \eta - 3, |w| - \eta - 2\}$ of the positions in w corresponding to the two characters following the prefix P of length $|w| - \eta - 4$. This means that such positions are exactly the positions corresponding to the end of each run in the output of $\texttt{bwt}(w)$. By Theorem 20 and by Theorem 6, each factor u in w has a circular occurrence in w crossing the position of Γ_2. Such an occurrence could not be entirely contained in w. In order to prove that Γ_2 is a string attractor of w, we have to show that w also admits an occurrence of the factor u crossing the positions of Γ_2. If $|u| \leq |P| + 1$, then its circular occurrence crossing a position in Γ_2 is entirely contained in w. Let us suppose $|u| \geq |P|+2$. If u is entirely contained in $\pi(w)$ and $|P|+2$ is a period of $\pi(w)$, then there exists an occurrence of u crossing the position $|P| + 1$ or $|P| + 2$.

Let us suppose u is not entirely contained in $\pi(w)$. If u doesn't cross the positions $|P|+1$ or $|P|+2$ then $u = xba$ or $u = xb$, where x is suffix of Q. It means that there exists an occurrence of u entirely contained in $\pi(w)$, so the thesis follows. Now, we suppose $x = a$. Then, $w = QabPba = PbaQba$. In this case $aQabPb$ and $bPbaQa$ are the smallest and the greatest conjugates in the lexicographic order, respectively. In this case the ending positions of each run in $\texttt{bwt}(w)$ correspond to the two characters following the prefix Q. So, we consider $\Gamma_1 = \{\eta+1, \eta+2\}$ and we prove that it is a string attractor for w. Let us consider a factor u of w. If $|u| \leq |P|+3$ then each circular occurrence of u crossing a position of Γ_1 is entirely contained in w. Let us consider $|u| \geq |P|+4$. If u is not entirely contained in $\pi(w)$ then it crosses at least a position in Γ_1. If u is entirely contained in $\pi(w)$, then, since $|P|+2$ is a period of $\pi(w)$ and $\pi(w)$ is palindrome, there exists an occurrence of u crossing a position in Γ_1.

The case $w = \pi(w)ab$ can be proved analogously by considering the starting characters of each run in the clustered output of $\texttt{bwt}(w)$. □

Example 23. *Given $w = ababaababaababababa \in Stand$, the PER-decompositions of w are $ababaababa.ab.aba.ba = aba.ba.ababaababa.ba$, then $\{11, 12\}$ is a (smallest) string attractor for w, since $\eta = 10$.*

Given $v = abaababaababa$, its PER-decompositions are $abaaba.ba.aba.ba = aba.ab.abaaba.ba$ and $\eta = 6$. So, $\{4, 5\}$ is a smallest string attractor for v.

Given $s = aaaaaa.ba.aaaaa.ab$, we can deduce that $\{7, 8\}$ is a string attractor for s. Let us consider the word $t = aa.ba.aaaaa.ab.aaaa$, that is a conjugate of s. One can check that, even if we can find a smallest string attractor $\Gamma^*(t) = \{3, 10\}$ of size 2, there is not any string attractor for t containing two consecutive positions.

The previous theorem shows an infinite family of finite binary words such that the size of the smallest string attractor is minimum. We remark that, in general, this is not the only family of binary words having a smallest string attractor with size 2, as shown in the following example.

Example 24. *A possible set of smallest string attractor for the words $u = a^n b^m$ or $w = b^n a^m$ is $\{n, n+1\}$. Moreover, words of the form $u = (ab)^{n_1}(ba)^{n_2}$ or $v = (ba)^{n_1}(ab)^{n_2}$ have a smallest string attractor of the form $\{2n_1, 2n_1 + 2n_2\}$.*

An open question is to characterize all the binary words with a string attractor of size 2. The question of characterizing all the words with string

attractors having size equal to the cardinality of the alphabet is also open for alphabets with more than two letters. Some examples of infinite families of words over an alphabet with cardinality greater than 2, with minimum size string attractors can be found in [27].

The large number of combinatorial properties of standard Sturmian words and their peculiarity of appearing as extremal cases for several string algorithms, makes it interesting to explore how some functions that measure the repetitiveness of strings and their compressibility behave if applied to standard Sturmian words. As shown in previous example, note that neither the size of the smallest string attractor nor its structure (two consecutive positions) are able, by themselves, to characterize the family of standard Sturmian words. The problem of characterizing standard Sturmian words by using string attractors is studied in next subsection.

5.2. Characterizing Standard Sturmian words via circular string attractors

In previous subsection we proved that for standard Sturmian words one can always find smallest string attractors whose positions are consecutive. Such a particular structure of the string attractor leads to investigate whether it can characterize some combinatorial properties of binary words. In this subsection we prove that the structure of the string attractor of a binary word is closely related to the (circular) factor complexity of the word itself. Furthermore, Example 23 shows that there exist conjugates of standard Sturmian words such that none of their smallest string attractors has two consecutive positions. In this subsection we provide a new characterization of the conjugacy classes of the standard Sturmian words by using a circular variant of string attractor. In particular, we introduce the new notion of *circular string attractor* which we use in this subsection as an investigative tool for the conjugacy classes of standard Sturmian words, but it could have an independent interest.

Definition 25. Let $w \in \Sigma^*$ and $n = |w|$. A set of γ_c positions $\Gamma_c = \{j_1, j_2, \dots j_{\gamma_c}\} \subseteq [1, n]$ is a *circular string attractor* of a word w if each circular factor of w has at least a circular occurrence that crosses a position of Γ_c. Moreover, we denote with γ_c^* the size of the smallest circular string attractor.

Before stating the main result of this subsection (Theorem 34), we show some combinatorial properties of circular string attractors. Despite the sim-

ilar definition, the concept of string attractor and circular string attractor can be considered as independent.

Example 26. *Let $w = a\underline{b}bb\underline{c}aa\underline{a}caaa$ be a word over the alphabet $\Sigma = \{a, b, c\}$. The set $\Gamma = \{2, 5, 8\}$ is a string attractor for w, since it covers any of its factors, but it is not a circular string attractor since the circular factor (in blue) caaaa escapes from it.*

On the other hand, the set $\Gamma_c = \{1, 4, 9\}$ is a circular string attractor for $w = \underline{a}bb\underline{b}c\underline{a}aa\underline{c}aaa$ but it is not a string attractor. In fact, the factor aaa (in blue), fully contained in w, is covered only if we consider its circular occurrence.

Although γ^* could arbitrarily increase when a conjugate of a word is considered (see Corollary 19), the behaviour of γ_c^* is different, as it can be seen in the following statement easily inferred from the definition.

Proposition 27. *Let w, w' be two words of the same conjugacy class. Then, $\gamma_c^*(w) = \gamma_c^*(w')$.*

In the following proposition we derive an upper bound on γ_c^* by considering the minimum size of the smallest string attractor of the words in the conjugacy class.

Proposition 28. *Let w a word in Σ^*. Then, $\gamma_c^*(w) \leq \gamma^*(v) + 1$ for each v conjugate of w.*

Proof. Let $\Gamma^*(v)$ be a string attractor having minimum size for v. If $\Gamma^*(v)$ is also a circular string attractor then we are done. Otherwise, there must be some *strictly* circular factor that overlaps two consecutive occurrences of w which is not fully contained in w and that is not covered by any position $j \in \Gamma^*(v)$. Note that these factors must cross the end and the beginning of two occurrences of w. Hence, the set $\Gamma^*(v) \cup \{1\}$ and $\Gamma^*(v) \cup \{n\}$ are both circular string attractors of w. □

The following corollary shows that the size of a smallest circular string attractor of a word can be arbitrarily lower than the size of a smallest string attractor of the word itself.

Corollary 29. *For each $t > 0$, there exists an alphabet Σ_t and a word $w_t \in \Sigma_t^*$, such that $\gamma^*(w_t) - \gamma_c^*(w_t) > t$.*

PROOF. Consider the word $w_t = v_1v_2v_3v_4v_5$ defined in the proof of the Proposition 18 and its conjugate $w'_t = v_2v_3v_4v_5v_1$. By using the same arguments as in Proposition 18 and by Proposition 28, it follows that $\gamma^*(w_t) - \gamma^*_c(w_t) \geq \gamma^*(w_t) - \gamma^*(w'_t) - 1 = (2t+3) - 1 = 2t + 2 > t$. □

It is easy to verify that the notion of circular string attractor of a word w is strictly related to the notion of string attractor of w^3, as proved in the following

Lemma 30. *Let $w \in \Sigma^n$ and $\Gamma_c = \{j_1, j_2, \dots j_{\gamma_c}\} \subseteq [1, n]$ a set of positions in w. Then, Γ_c is a circular string attractor of w if and only if $\Gamma' = \{j_k + n \mid j_k \in \Gamma_c\}$ is a string attractor of w^3.*

PROOF. ($\Longrightarrow$) Note that the positions in Γ' correspond to the positions of Γ_c in the central occurrence of w. Moreover, it is easy to check that any k-length factor u of w^3 is a circular factor of w, with $k = 1, 2 \dots, n$. In fact, if u lies entirely in an occurrence of w, then it is a circular factor of w as well. Otherwise, u overlaps two consecutive occurrences of w, which is still a circular factor of w. Moreover, as we have picked the positions in Γ', every factor u has an occurrence crossing a position $j_i \in \Gamma'$, either this occurrence is fully contained in w or lies across two consecutive w's. For any factor v such that $|v| > n$, if $v = w[i, n] \cdot w \cdot w[1, j]$, with $1 \leq i, j \leq n$, then v has an occurrence which crosses all the positions in Γ', since it contains the central occurrence of w in w^3. Otherwise, $v = w[i, n] \cdot w[1, j]$, with $|w[i, n]| + |w[1, j]| > n$. Since v appears twice in w^3 and w^3 has period n, one of the two occurrences of v has to cross a position $j_i \in \Gamma'$ in the central occurrence of w.

($\Longleftarrow$) Since Γ' is a string attractor for w^3 and, as mentioned above, any k-length factor u in w^3 is a circular factor of w with $k = 1, 2, \dots, n$, using the previous reasoning we can easily verify that Γ_c is a circular string attractor for w. □

Example 31. *Let $w = aa\underline{b}a\underline{a}caab\underline{c}$ be a word on the alphabet $\Sigma = \{a, b, c\}$. A circular string attractor of the word w is $\Gamma_c = \{3, 5, 10\}$. In fact, every factor fully contained in w has an occurrence crossing a position $j \in \Gamma' = \{13, 15, 20\}$, except for $u = caab$ and its prefixes ca and caa. However, u also occurs between two consecutive occurrences of w:*

$$w^3 = aabaacaabc \cdot aa\underline{b}a\underline{a}caab\underline{c} \cdot aabaacaabc.$$

As we can see, u and its prefixes cross the position $20 \in \Gamma'$. For what concerns any other factor that lies in $w \cdot w$ but does not appear in w, note that the position 20 covers them all too.

Remark 32. *Due to the distribution of the circular factors in w, w^2 may not be sufficient to our purpose. For instance, consider the word $w = aa\underline{b}ab\underline{a}a$ that admits the set $\Gamma_c^* = \{3, 6\}$ as smallest circular string attractor (the underlined positions). Note that, if one would consider the word $w^2 = aa\underline{b}ab\underline{a}a.aa\overline{b}ab\overline{a}a$, neither $\Delta_1 = \{3, 6\}$ (underlined positions) nor $\Delta_2 = \{10, 13\}$ (overlined positions) are string attractors for w^2, since the factors $aaab$ and $baaa$ escape from Δ_1 and Δ_2 respectively. On the other hand, picking the positions of Γ_c^* in the central occurrence of w in w^3 covers them both*

$$w^3 = aababa\underline{a}.aa\underline{b}ab\underline{a}a.\underline{a}ababaa.$$

Clearly, the same argument holds for any w^n with $n \geq 3$ and picking the positions of Γ_c^* within any internal occurrence of w (i.e. every occurrence of w except the first and the last).

In the following lemma we consider a binary word $w = a_1 a_2 \dots a_n$ that admits a smallest circular string attractor $\Gamma_c^* = \{i, j\}$, with $i < j$. Let $d = \min\{j - i, n - j + i\}$ be the *distance* between the positions i and j. Note that $d \leq \frac{n}{2}$. If $d = 1$ the two positions i and j are called *consecutive*. Set $d' = n - d = \max\{j - i, n - j + i\}$. One has $1 \leq d \leq \frac{n}{2} \leq d' \leq n - 1$. The next lemma provides an upper bound for the circular factor complexity of the word w. Since $f_w(k) \leq c_w(k)$, for each $k \geq 1$, an analogous bound for f_w also holds.

Lemma 33. *Let $w = a_1 a_2 \dots a_n$ be a binary word that admits a smallest circular string attractor $\Gamma_c^* = \{i, j\}$ consisting of two positions at distance d. Then, for $1 \leq k \leq n - 1$, we have:*

$$c_w(k) \leq \begin{cases} 2k, & \text{if } k \leq d \\ k + d, & \text{if } d < k \leq d' \\ n & \text{if } k > d' \end{cases}$$

Proof. Let us distinguish the three cases:

- ($k \leq d$) Since a position is crossed by at most k distinct circular factors of length k and that any of these does not cross both i and j, we have that the number of distinct k-length circular factors in w is bounded by $|\Gamma_c^*|k = 2k$.

- ($d < k \leq d'$) In this case, there are some factors of size k that cross both the positions i and j. Let us suppose that $d = j - i$. Then, we can count these factors by sliding a k-length circular window from the factor ending in j to the one starting in i. Since i and j are at distance d, these factors are $k - d$. Thus, the bound on the number of distinct factors for $d < k \leq d'$ is given by $2k - (k - d) = k + d$. The case $d = n - j + 1$ can be treated analogously by sliding the circular window from the circular factor ending in i to the one starting in j, obtaining the same bound.

- ($k > d'$) Like the previous case, we have to further remove the other circular factors that are longer than the maximum distance d' and cross both positions i and j (this time from the other side). Hence, these factors are $2k - (k - d) - (k - d') = d + d' = n$. □

By refining a result stated in [27], we show that the notion of circular string attractors can be used to state the main result of this subsection providing a new characterization of the conjugacy classes of the standard Sturmian words.

Theorem 34. *Let w be a primitive word. The word w is a conjugate of a standard Sturmian word if and only if w admits a smallest circular string attractor consisting of two consecutive positions.*

In the proof of the Theorem 34 we use the following result, proved in [5], that states that the conjugates of standard Sturmian words are uniquely characterized by the circular factor complexity.

Theorem 35 ([5]). *Let w be a word of length $n \geq 2$. The following statements are equivalent:*

1. *w is conjugate of a standard Sturmian word;*

2. *for $k = 0, 1, \ldots, n - 1$, $c_w(k) = k + 1$;*

3. $c_w(n-2) = n-1$ *and* w *is primitive.*

PROOF (OF THEOREM 34). ($\Longrightarrow$) By combining the Theorem 20 and Theorem 6, we can find a circular string attractor by taking the position of the first or the last occurrence from each run of equal-letter in the $\texttt{bwt}(w)$. In particular, the two minimum circular string attractors obtained correspond to the string attractors built for standard Sturmian words in Theorem 22, which contain two consecutive positions.

($\Longleftarrow$) From Theorem [5, Lemma 4.1], we know that a word w is primitive if and only if $c_w(k) \geq k+1$, for $k = 1, \ldots, n-1$. Moreover, in Lemma 33 we have proved that a circular string attractor with two elements defines an upper-bound on the number of distinct circular factors of length k in w. Since the positions in the minimum circular string attractor are consecutive, $d = 1$ and $d' = n-1$. For $k = 1$ (i.e. the case $k \leq d$) we have $c_w(1) = 2$ distinct factors, which are the letters of the alphabet $\{a, b\}$. For $k = 2, \ldots, n-1$ (i.e. the case $d < k \leq d'$) we have $c_w(k) = k + d = k + 1$ distinct factors. Note that it does not exist a value of $k = 1, 2, \ldots, n-1$ greater than d', meaning that we can ignore the third case. Since lower and upper-bound overlap, we can stand that for any $k \in \{1, \ldots, n-1\}$ we have $c_w(k) = k+1$. Finally, using Theorem 35, we have that w is conjugate to a standard Sturmian word. □

Remark 36. *It is known that each conjugacy class of standard Sturmian words exactly contains two standard Sturmian words* Xab *and* Xba*, where* $X \in PER$*. This means that each word in the conjugacy class has two circular string attractors consisting of* 2 *consecutive positions, as depicted in Figure 2.*

6. String Attractors in other infinite families of words

String attractors have been introduced, in the field of Data Compression, as a measure of the compressibility of repetitive strings. The notion of repetitiveness can be described in various ways in Combinatorics on Words and it is however related to how frequently the factors of a word appear. In this section we describe the problem of finding string attractors for other well known infinite families of words, such as Thue-Morse words and de Bruijn words, and we compare the size of a smallest string attractor with the combinatorial measure δ that has been studied in [20] to overcome the problem of the

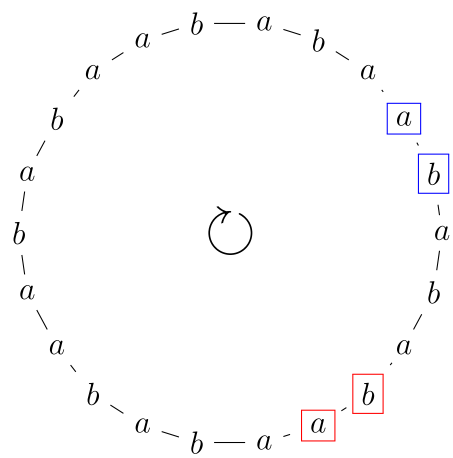

Figure 2: Circular representation of the standard Sturmian words $aba.ba.ababaababaababa.ab$ and $ababaababaababa.ab.aba.ba$. Both couples of boxed consecutive positions represent a circular string attractor for each word in the correspondent conjugacy class.

NP-hardness of the computation of γ^*. It is related to the factor complexity and defined as follows: given a word $w \in \Sigma^*$,

$$\delta(w) = \max\{f_w(k)/k, 1 \leq k \leq |w|\}.$$

It is known that δ can be computed in linear time and that $\delta(w) \leq \gamma^*(w)$ (this is a consequence of Proposition 9), but it seems that γ^* and δ are able to highlight different combinatorial properties of the words.

We remark that there are infinite families of words for which the values of both δ and γ^* are constant. For instance, if w is a standard Sturmian word we can easily deduce that $\delta(w) = 2$, by using combinatorial properties of such a word.

However, if we consider the word w (of length $n = 2^m$) over the alphabet $\{a, b\}$ having b's in positions 2^i for $0 \leq i \leq m$ and a's elsewhere, i.e.

$$w = bbabaaabaaaaaaab \cdots ba^{2^i-1}b \cdots ba^{2^{m-1}-1}b,$$

it is possible to verify that the value of δ is $O(1)$ whereas the value of γ^* is m, that is logarithmic on the length of w (cf. [20]).

6.1. String Attractors in Thue-Morse Words

In this subsection we give some details on the problem of finding a smallest string attractor for the family of finite binary Thue-Morse words. Thue-Morse words are a sequence of words obtained by the iterated application of a morphism as described below.

Definition 37. Let us consider the alphabet $\Sigma = \{a, b\}$ and the morphism $\varphi : \Sigma^* \mapsto \Sigma^*$ such that $\varphi(a) = ab$ and $\varphi(b) = ba$. Let us denote by $t_n = \varphi^n(a)$ the n-th iterate of the morphism φ that is called the n-th Thue-Morse word.

Note that at each iteration of φ the length of the word is doubled, therefore the n-th Thue-Morse word has length 2^n. The n-th Thue-Morse words for $n = 3, 4, 5$ are shown in Figure 3.

By using a result in [7] on the enumeration of factors in Thue-Morse words one can deduce that $\delta(t_n) = O(1)$. In particular, in [21] it has been proved that $\delta(t_n) = \frac{10}{3+2^{4-n}}$ for $n \geq 3$. So, the following lower bound for γ^* can be deduced. Note that the value for $n = 3$ is found by using an exhaustive search.

Proposition 38. *Let $t_n = \varphi^n(a)$ be the n-th Thue-Morse word with $n \geq 3$. Then $\gamma^*(t_n) \geq 3$.*

The problem of computing the smallest string attractor for finite Thue-Morse words has been addressed in [27] where it was conjectured that the size of the smallest string attractor is logarithmic with respect to the length of the word, i.e. $\gamma^*(t_n) = n$. In [21] such a conjecture has been disproved. In fact, it has been shown that $\gamma^*(t_n) = 4$ for $n \geq 4$ and an explicit nice construction of a smallest string attractor for t_n is given.

Theorem 39 (Theorems 2 and 3 in [21]). *For any $n \geq 4$, the set*

$$K_n = \{2^{n-2}, 3 \cdot 2^{n-3}, 2^{n-1}, 3 \cdot 2^{n-2}\}$$

is a smallest string attractor of t_n, then $\gamma^*(t_n) = 4$.

String attractors for t_n, with $n = 3, 4, 5$, are shown in Fig. 3.

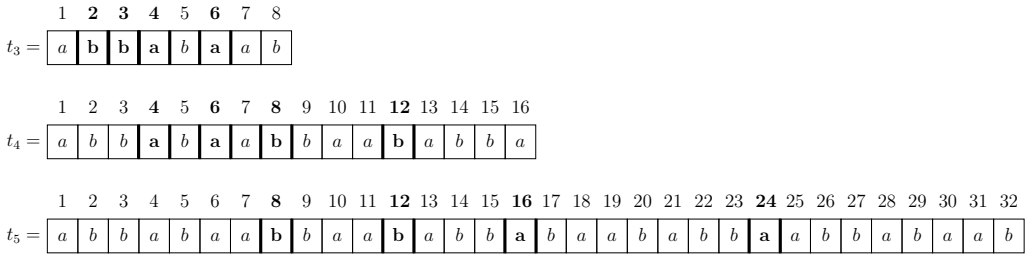

Figure 3: String attractor K_n for the word $t_n = \varphi^n(a)$, with $n = 3, 4, 5$ (the positions in K_n are in bold), i.e. $K_3 = \{2, 3, 4, 6\}$, $K_4 = \{4, 6, 8, 12\}$, $K_5 = \{8, 12, 16, 24\}$. Note that K_3 is not a smallest string attractor for t_3 since the set $\{3, 5, 6\}$ is a string attractor too.

6.2. Attractors in de Bruijn words

A *de Bruijn* sequence (or word) B of order k on an alphabet Σ of size σ, is a circular sequence in which every possible length-k string on Σ occurs exactly once as a substring.

De Bruijn words are widely studied in Combinatorics on Words, and all of them can be constructed by considering all the Eulerian walks on de Bruijn graphs. All the de Bruijn sequences of order k over an alphabet of size σ have length σ^k. For instance the (circular) word $w = aaaababbbbabaabb$ is a de Bruijn word of order 4 over the alphabet $\{a, b\}$. In fact one can verify that all strings of length 4 over $\{a, b\}$ appear as factor of w just once.

Since we are here interested to linear and not to circular words, it is easy to verify that in order to have linear words containing all the k-length factors exactly once, it is sufficient to consider any linearization of the circular de Bruijn word of order k (that is, we cut the circular word in any position to get a linear word) and concatenate it with a word equal to its own prefix of length $k-1$. Therefore its length is $\sigma^k + k - 1$. We call such words *linear de Bruijn sequences (or words)*. For instance the linear de Bruijn word corresponding to the circular one in the above example is the word $w' = aaaababbbbabaabbaaa$ of length of $2^k + k - 1$. In [22] the following theorem is proved.

Theorem 40. *The number of phrases $c(n)$ in a LZ-parsing of a sequence of length n over an alphabet of size σ satisfies:*

$$c(n) < \frac{n}{(1 - \epsilon_n) \log_\sigma n}$$

where $\epsilon_n = 2^{\frac{1+\log_\sigma(\log_\sigma(\sigma n))}{\log_\sigma n}}$.

When a linear de Bruijn word B of order k is considered, by combining Theorem 40, Theorem 8, Proposition 9 and the fact that the prefix and the suffix of B of length $k-1$ are equal, we get the following upper and lower bounds for a smallest string attractor of B.

Proposition 41. *Let B be a linear de Bruijn sequence of order k and length $n+k-1$ over an alphabet of size σ ($n = \sigma^k$). Then the cardinality γ^* of a smallest string attractor for B satisfies:*

$$\frac{n}{\log_\sigma n} \leq \gamma^* < \frac{n}{(1-\epsilon_n)\log_\sigma n}$$

where $\epsilon_n = 2\frac{1+\log_\sigma(\log_\sigma(\sigma n))}{\log_\sigma n}$.

This means that γ^* for a linear de Bruijn word of length n grows asymptotically as $\frac{n}{\log n}$, corresponding to the worst case for the size of a smallest string attractor of any word over the constant alphabet Σ. Notice that the lower bound is somehow intuitively expected, since all the words of length k appear only once in B, therefore two consecutive positions in any string attractor cannot be farther than k. Moreover, one can easily verify that $\delta(B) = \frac{n}{\log n}$. For instance a smallest string attractor for $w' = aaaababbbbabaabbaaa$ is $\{4, 8, 12, 16\}$.

7. Conclusion and Open Problems

In this paper we have studied the notion of string attractor from a combinatorial point of view. In particular, we have given an explicit construction of a smallest string attractor for the well known infinite family of standard Sturmian words. By using their combinatorial properties, the construction provides a smallest string attractor whose size is 2, that is the minimum possible string attractor for two-letters alphabets. It is open the question to characterize all those words whose smallest string attractor has size equal to the cardinality of the alphabet.

We have introduced the new notion of circular string attractor that allows to uniquely characterize the conjugates of standard Sturmian words, in the sense that a word has a circular string attractor having two consecutive positions if and only if it is a conjugate a standard Sturmian word. It would be interesting to investigate whether the size and the structure of a (circular) string attractor could uniquely characterize other infinite families of words.

Other variants of the notion of string attractor have been studied [18], such as k-attractor and k-sharp attractor in which it is required that their positions must cross the occurrences of all the distinct factors of length at most k or exactly equal to k, respectively. It would be interesting to explore whether such notions can be used to investigate other combinatorial properties of words.

Acknowledgements

The authors are very grateful to the anonymous referees for their constructive comments and suggestions that have improved the quality of the manuscript.

S. Mantaci, G. Rosone and M. Sciortino are partially supported by the project MIUR-SIR CMACBioSeq ("Combinatorial methods for analysis and compression of biological sequences") grant n. RBSI146R5L.

References

[1] Baturo, P., Rytter, W.: Compressed string-matching in standard Sturmian words. Theoret. Comput. Sci. **410**(30), 2804 – 2810 (2009)

[2] Berstel, J., de Luca, A.: Sturmian words, Lyndon words and trees. Theoret. Comput. Sci. **178**(1), 171 – 203 (1997)

[3] Bonomo, S., Mantaci, S., Restivo, A., Rosone, G., Sciortino, M.: Suffixes, Conjugates and Lyndon words. In: DLT, Lect. Notes Comput. Sc., vol. 7907, pp. 131–142. Springer (2013)

[4] Bonomo, S., Mantaci, S., Restivo, A., Rosone, G., Sciortino, M.: Sorting conjugates and suffixes of words in a multiset. Int. J. Found. Comput. S. **25**(8), 1161–1175 (2014)

[5] Borel, J.P., Reutenauer, C.: On Christoffel classes. RAIRO-Theor. Inf. Appl. **40**(1), 1527 (2006)

[6] Boucher, C., Gagie, T., Kuhnle, A., Langmead, B., Manzini, G., Mun, T.: Prefix-free parsing for building big BWTs. Algorithms Mol. Biol. **14**(1), 13:1–13:15 (2019)

[7] Brlek, S.: Enumeration of factors in the Thue-Morse word. Discrete Appl. Math. **24**(1-3), 83–96 (1989)

[8] Burrows, M., Wheeler, D.J.: A block sorting data compression algorithm. Tech. rep., DIGITAL System Research Center (1994)

[9] Castiglione, G., Restivo, A., Sciortino, M.: Circular Sturmian words and Hopcroft's algorithm. Theoret. Comput. Sci. **410**(43), 4372–4381 (2009)

[10] Castiglione, G., Restivo, A., Sciortino, M.: On extremal cases of Hopcroft's algorithm. Theoret. Comput. Sci. **411**(38-39), 3414–3422 (2010)

[11] Castiglione, G., Restivo, A., Sciortino, M.: Hopcroft's Algorithm and Cyclic Automata. In: LATA. Lect. Notes Comput. Sc., vol. 5196, pp. 172–183. Springer (2008)

[12] Gagie, T., Navarro, G., Prezza, N.: Optimal-time text indexing in bwt-runs bounded space. In: SODA. pp. 1459–1477. SIAM (2018)

[13] Gagie, T., Navarro, G., Prezza, N.: Fully functional suffix trees and optimal text searching in BWT-runs bounded space. J. ACM **67**(1) (2020)

[14] Giancarlo, R., Restivo, A., Sciortino, M.: From first principles to the Burrows and Wheeler transform and beyond, via combinatorial optimization. Theoret. Comput. Sci. **387**(3), 236 – 248 (2007)

[15] Guerrini, V., Louza, F., Rosone, G.: Metagenomic analysis through the extended Burrows-Wheeler transform. BMC Bioinformatics **21** (2020)

[16] Kempa, D., Prezza, N.: At the roots of dictionary compression: string attractors. In: STOC 2018. pp. 827–840. ACM (2018)

[17] Kempa, D., Kociumaka, T.: Resolution of the Burrows-Wheeler Transform conjecture. CoRR **abs/1910.10631** (2019), accepted to the 61st Annual Symposium on Foundations of Computer Science (FOCS 2020)

[18] Kempa, D., Policriti, A., Prezza, N., Rotenberg, E.: String Attractors: Verification and Optimization. In: ESA. Leibniz International Proceedings in Informatics (LIPIcs), vol. 112, pp. 52:1–52:13. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2018)

[19] Knuth, D., Morris, Jr., J., Pratt, V.: Fast pattern matching in strings. SIAM J. Comput. **6**(2), 323–350 (1977)

[20] Kociumaka, T., Navarro, G., Prezza, N.: Towards a definitive measure of repetitiveness. CoRR **abs/1910.02151** (2019), accepted to the 14th Latin American Symposium on Theoretical Informatics (LATIN 2020)

[21] Kutsukake, K., Matsumoto, T., Nakashima, Y., Inenaga, S., Bannai, H., Takeda, M.: On repetitiveness measures of Thue-Morse words. In: SPIRE. Lect. Notes Comput. Sc., vol. 12303, pp. 213–220. Springer (2020)

[22] Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE T. Inform. Theory **22**(1), 75–81 (1976)

[23] Lothaire, M.: Algebraic Combinatorics on Words. Cambridge University Press (2002)

[24] Louza, F.A., Telles, G.P., Gog, S., Zhao, L.: Computing Burrows-Wheeler Similarity Distributions for String Collections. In: SPIRE 2018. Lect. Notes Comput. Sc., vol. 11147, pp. 285–296. Springer (2018)

[25] de Luca, A., Mignosi, F.: Some combinatorial properties of sturmian words. Theoret. Comput. Sci. **136**(2), 361–385 (1994)

[26] de Luca, A.: Sturmian words: Structure, combinatorics, and their arithmetics. Theor. Comput. Sci. **183**(1), 45–82 (1997)

[27] Mantaci, S., Restivo, A., Romana, G., Rosone, G., Sciortino, M.: String attractors and combinatorics on words. In: ICTCS. CEUR Workshop Proceedings, vol. 2504, pp. 57–71. CEUR-WS.org (2019)

[28] Mantaci, S., Restivo, A., Rosone, G., Sciortino, M.: Suffix array and Lyndon factorization of a text. Journal of Discrete Algorithms **28**, 2–8 (2014)

[29] Mantaci, S., Restivo, A., Rosone, G., Sciortino, M., Versari, L.: Measuring the clustering effect of BWT via RLE. Theoret. Comput. Sci. **698**, 79–87 (2017)

[30] Mantaci, S., Restivo, A., Sciortino, M.: Burrows-Wheeler transform and Sturmian words. Inform. Process. Lett. **86**, 241–246 (2003)

[31] Policriti, A., Prezza, N.: LZ77 Computation Based on the Run-Length Encoded BWT. Algorithmica **80**(7), 1986–2011 (2018)

[32] Prezza, N.: String attractors. CoRR **abs/1709.05314** (2017)

[33] Prezza, N., Pisanti, N., Sciortino, M., Rosone, G.: SNPs detection by eBWT positional clustering. Algorithm. Mol. Biol. **14**(1), 3:1–3:13 (2019)

[34] Prezza, N., Pisanti, N., Sciortino, M., Rosone, G.: Variable-order reference-free variant discovery with the Burrows-Wheeler transform. BMC Bioinformatics **21** (2020)

[35] Restivo, A., Rosone, G.: Balancing and clustering of words in the Burrows-Wheeler transform. Theoret. Comput. Sci. **412**(27), 3019 – 3032 (2011)

[36] Sciortino, M., Zamboni, L.Q.: Suffix automata and standard sturmian words. In: DLT. Lect. Notes Comput. Sc., vol. 4588, pp. 382–398. Springer (2007)

[37] Ziv, J., Lempel, A.: A universal algorithm for sequential data compression. IEEE T. Inform. Theory **23**(3), 337–343 (1977)

[38] Ziv, J., Lempel, A.: Compression of individual sequences via variable-length coding. IEEE T. Inform. Theory **24**, 530–536 (1978)