Smallest Suffixient Sets as a Repetitiveness Measure

Gonzalo Navarro[1,2], Giuseppe Romana[3], and Cristian Urbina[1,2]

[1] Department of Computer Science, University of Chile, Chile
[2] Center for Biotechnology and Bioengineering (CeBiB), Chile
[3] Department of Mathematics and Computer Science, University of Palermo, Italy
gnavarro@uchile.cl,giuseppe.romana01@unipa.it,crurbina@dcc.uchile.cl

Abstract. Suffixient sets are a novel combinatorial object that capture the essential information of repetitive strings in a way that, provided with a random-access mechanism, supports various forms of pattern matching. In this paper we study the size χ of the smallest suffixient set as a repetitiveness measure: we place it between known measures and study its sensitivity to various string operations.

Keywords: Repetitive sequences · Text compressibility · Burrows-Wheeler Transform.

1 Introduction

The study of repetitive string collections has recently attracted considerable interest from the stringology community, triggered by practical challenges such as representing huge collections of similar strings in a way that they can be searched and mined directly in highly compressed form [24,25]. An example is the *European '1+ Million Genomes' Initiative*[4], which aims at sequencing over a million human genomes: while this data requires around 750TB of storage in raw form (using 2 bits per base), the high similarity between human genomes would allow storing it in querieable form using two orders of magnitude less space.

An important aspect of this research is to understand how to measure repetitiveness, especially when those measures reflect the size of compressed representations that offer different access and search functionalities on the collection. Various repetitiveness measures have been proposed, from abstract lower bounds to those related to specific text compressors and indices; a relatively up-to-date survey is maintained [26]. Understanding how those measures relate to each other sheds light on what search functionality is obtained at what space cost.

A relevant measure proposed recently is the size χ of the smallest *suffixient set* of the text collection [6], whose precise definition will be given later. Within $O(\chi)$ size, plus a random-access mechanism on the string, it is possible to support some text search functionalities, such as finding one occurrence of a pattern, or

[4] https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes

finding its maximal exact matches (MEMs), which is of central use on various bioinformatic applications [4].

While there has been some work already on how to build minimal suffixient sets and how to index and search a string within their size, less is known about that size, χ, as a measure of repetitiveness. It is only known [6] that $\gamma = O(\chi)$ and $\chi = O(\overline{r})$ on every string family, where γ is the size of the smallest *string attractor* of the collection (a measure that lower bounds most repetitiveness measures) [18] and $\overline{r}$ is the number of equal-letter runs of the Burrows-Wheeler Transform (BWT) [3] of the reversed string.

In this paper we better characterize χ as a repetitiveness measure. First, we study how it behaves when the string undergoes updates, showing in particular that it grows by $O(1)$ when appending or prepending symbols, but that it can grow by $\Omega(\log n)$ upon arbitrary edit operations or rotations, and by $\Omega(\sqrt{n})$ when reversing the string. Second, we show that $\chi = O(r)$ on every string family, where r is the number of equal-letter runs of the BWT of the string. We also show that there are string families where $\chi = o(v)$, where v is the size of the smallest lexicographic parse [27] (an alternative to the size of the Lempel-Ziv parse [20], which behaves similarly). In particular, this holds on the Fibonacci strings, where we fully characterize the only 2 smallest suffixient sets of size 4, and further prove that $\chi \leq \sigma + 2$ on all substrings of episturmian words over an alphabet of size σ. Since $v = O(r)$ on all string families, this settles χ as a strictly smaller measure than r, which is a more natural characterization than in terms of the reverse string. We also show that χ is incomparable with most "copy-paste" based measures [24], as there are families where it is strictly smaller and others where it is strictly larger than any of those measures.

This result relates to the important question of whether a measure μ is *reachable* (i.e., one can represent the string within $O(\mu)$ space), *accessible* (i.e., one can access any string position from an $O(\mu)$-size representation, in sublinear time), or *searchable* (i.e., one can search for patterns in sublinear time within space $O(\mu)$). Measure r is, curiously, the only one to date being searchable but unknown to be accessible. Now χ emerges as a measure smaller than r, which can search if provided with a mechanism to efficiently access substrings (r does not need access to support searches). Unlike r, χ is yet unknown to be reachable (as its relation to the smallest known reachable measure, the size b of the smallest bidirectional macro scheme [30], remains unknown). As said, it is known that $\gamma = O(\chi)$, but γ is also unknown to be reachable.

2 Preliminaries

An *ordered alphabet* $\Sigma = \{a_1, \ldots, a_\sigma\}$ is a finite set of symbols equipped with a total order $<$ such that $a_1 < a_2 < \cdots < a_\sigma$. When $\sigma = 2$, we assume $\Sigma = \{\mathtt{a}, \mathtt{b}\}$ with $\mathtt{a} < \mathtt{b}$, and define the *complement* as $\overline{\mathtt{a}} = \mathtt{b}$ and $\overline{\mathtt{b}} = \mathtt{a}$. The special symbol $\$$, if appears, is always assumed to be the smallest on the alphabet.

A *string* $w[1\,..\,n]$ (or simply w if it is clear from the context) of *length* $|w| = n$ over the alphabet Σ is a sequence $w[1]w[2]\ldots w[n]$ of symbols where $w[i] \in \Sigma$

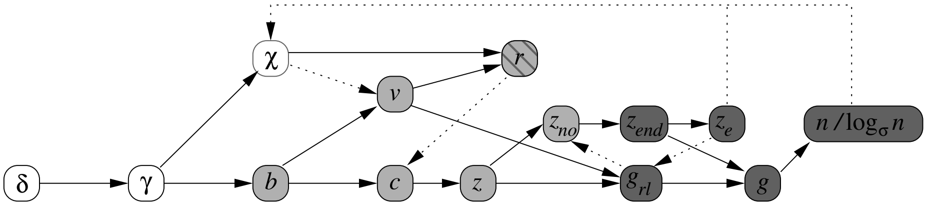

Fig. 1. Relations between relevant repetitiveness measures and how our results place χ among them. An arrow $\mu_1 \to \mu_2$ means that $\mu_1 = O(\mu_2)$ for all strings and, save for $c \to z$, there is a string family where $\mu_1 = o(\mu_2)$. The dotted arrows mark only this last condition, so they are not transitive. Measures in light gray nodes are known to be reachable; those in dark gray are accessible and searchable; and r is hatched because it is searchable but unknown to be accessible.

for all $i \in [1, n]$. The *empty string* string of length 0 is denoted ϵ. We denote by Σ^* the set of all strings over Σ. Additionally, we let $\Sigma^+ = \Sigma^* \setminus \{\epsilon\}$ and $\Sigma^k = \{w \in \Sigma^* \mid |w| = k\}$. We denote by $w[i\,..\,j]$ the subsequence $w[i]w[i+1]\dots w[j]$. If $x[1\,..\,n]$ and $y[1\,..\,m]$ are strings, we define the *concatenation operation* applied on x and y, as the string obtained by juxtaposing these two strings, that is, $x \cdot y = x[1]x[2]\cdots x[n]y[1]\cdots y[m] = xy$. A string x is a *substring* of w if $w = yxz$ for some $y, z \in \Sigma^*$. A string x is a *prefix* of w if $w = xy$ for some $y \in \Sigma^*$. Analogously, x is a *suffix* of w if $w = yx$ for some $y \in \Sigma^*$. We say that substrings, prefixes, and suffixes are *non-trivial* if they are different from w and ϵ. The set of substrings of w is denoted by $\mathcal{F}_w$. We also let $\mathcal{F}_w(k) = \mathcal{F}_w \cap \Sigma^k$. The *reverse* of a finite string w is the string $w^R = w[n] \cdot w[n-1] \cdots w[1]$. We denote by $\mathcal{R}(w)$ the multiset of rotations of $w[1\,..\,n]$, that is, $\mathcal{R}(w) = \{w[i+1\,..\,n]w[1\,..\,i] \mid i \in [1\,..\,n]\}$. Moreover, we let $\mathcal{R}_x(w)$ be the multiset of rotations of w prefixed by the string x. The *Burrows-Wheeler transform* (BWT) of a string w, denoted $\texttt{BWT}(w)$, is the transformation of w obtained by collecting the last symbol of all rotations in $\mathcal{R}(w)$ in lexicographic order. The *BWT matrix* $B(w)$ of w is the $(n \times n)$-matrix where the i-th row is the i-th rotation of w in lexicographic order.

A *right-infinite string* $\mathbf{w}$ —we use **boldface** to emphasize its infinite length— over Σ is any infinite sequence $\mathbb{Z}^+ \to \Sigma$. The set of all infinite strings over Σ is denoted Σ^ω. A substring of $\mathbf{w}$ is the finite string $\mathbf{w}[i\,..\,j]$ for any $1 \le i \le j$. A prefix of $\mathbf{w}$ is a finite substring of the form $\mathbf{w}[1\,..\,n]$ for some $n \ge 0$. The *substring complexity function* $P_{\mathbf{w}}(k) : \mathbb{Z}^+ \cup \{0\} \to \mathbb{Z}^+$ counts the number of distinct substrings of length k in $\mathbf{w}$, for any $k \in \mathbb{Z}^+ \cup \{0\}$, that is, $P_{\mathbf{w}}(k) = |\mathcal{F}_{\mathbf{w}}(k)|$. For a finite string $w[1\,..\,n]$, the domain of P_w is restricted to $[0\,..\,n]$.

2.1 Measures of repetitiveness

In this work, we will relate χ, in asymptotic terms, with several well-established measures of repetitiveness [24,26]: $\delta = \max_{k\in[0\,..\,n]}(\mathcal{F}_w(k)/k)$ (a measure of string complexity), γ (the smallest string attractor), b (the size of the smallest bidirectional macro scheme), z (the size of a Lempel-Ziv parse), z_{no} (the same without allowing phrases to overlap their sources), z_e (the size of the greedy LZ-End parse), z_{end} (the size of the minimal LZ-End parse), v (the size of the smallest lexicographic parse), r (the number of equal-letter runs in the BWT of the string), g (the size of the smallest context-free grammar generating only the string), g_{rl} (the same allowing run-length rules), and c (the size of the smallest collage system generating only the string). Except for δ, γ and r, these measures are said to be *copy-paste* because they refer to a way of cutting the sequence into chunks that can be copied from elsewhere in the same sequence. Indeed, δ and γ are lower-bound measures, the former known to be unreachable and the latter unknown to date to be reachable; all the others are. The smallest measures known to be accessible (and searchable) are z_{end} and g_{rl}, and r is searchable but unknown to be reachable.

The known relations between those measures are summarized in Fig. 1, where we have added the results we obtain in this paper with respect to χ.

2.2 Edit operations and sensitivity functions

The so-called *edit operations* are *insertion*, *substitution* and *deletion* of a single character on a string. We denote $\mathtt{ins}_\Sigma(w)$, $\mathtt{sub}_\Sigma(w)$, $\mathtt{del}_\Sigma(w)$ the sets of strings that can be obtained by applying an edit operation to w. In addition, we let $\mathtt{prepend}_\Sigma(w)$ and $\mathtt{append}_\Sigma(w)$ be $\mathtt{ins}_\Sigma(w)$ restricted to the insertion being made at the beginning and the end of the string, respectively.

A repetitiveness measure μ is *monotone* or *non-decreasing* to the insertion of a single character if $\mu(w') - \mu(w) \geq 0$ for any w and $w' \in \mathtt{ins}_\Sigma(w)$. More generally, the *additive sensitivity* and *multiplicative sensitivity* functions of a repetitiveness measure μ to the insertion of a single character are the maximum possible values of $\mu(w') - \mu(w)$ and $\mu(w')/\mu(w)$, respectively. We define the concept of monotonicity and sensitivity functions for the remaining string operations analogously.

3 Suffixient Sets and the Measure χ

In this section we define the central combinatorial objects and measures we analyse on this work. Note that some of our definitions are slightly different from their original formulation [4,5], because we do not always assume that all strings are $-terminated.

Definition 1 (Right-maximal substrings and right-extensions [4,5]). *Let $w \in \Sigma^*$. A substring x of w is* right-maximal *if there exist at least two distinct symbols $a, b \in \Sigma$ such that both xa and xb are substrings of w. For any right-maximal substring x of w, the substrings xa with $a \in \Sigma$ are called* right-extensions. *We denote the set of right-extensions in w by $E_r(w) = \{xa \mid \exists b : b \neq a, xa \in \mathcal{F}_w, xb \in \mathcal{F}_w\}$.*

We distinguish a special class of right-extensions that are not suffixes of any other right-extension.

Definition 2 (Super-maximal extensions [4,5]). *The set of* super-maximal extensions *of w is $\mathcal{S}_r(w) = \{x \in E_r(w) \mid \forall y \in E_r, y = zx \Rightarrow z = \varepsilon\}$. Moreover, we let $\texttt{sre}(w) = |\mathcal{S}_r(w)|$.*

We now define suffixient sets for strings not necessarily $\texttt{\$}$-terminated; we introduce later the special terminator $\texttt{\$}$.

Definition 3 (Suffixient set [4,5]). *Let $w[1\,..\,n] \in \Sigma^*$. A set $S \subseteq [1\,..\,n]$ is a* suffixient set *for w if for every right-extension $x \in E_r(w)$ there exists $j \in S$ such that x is a suffix of $w[1\,..\,j]$.*

Intuitively, a suffixient set is a collection of positions of $[1\,..\,|w|]$ capturing all the right-extensions appearing in w. The smallest suffixient sets, which are suffixient sets of minimum size, have also been characterized in terms of super-maximal right-extensions. The next definition simplifies the original one [4,5].

Definition 4 (Smallest suffixient set). *Let $w[1\,..\,n] \in \Sigma^*$. A suffixient set $S \subseteq [1\,..\,n]$ is a* smallest suffixient set *for w if there is a bijection $pos : \mathcal{S}_r \to S$ such that every $x \in \mathcal{S}_r$ is a suffix of $w[1\,..\,pos(x)]$.*

In its original formulation, the measure χ is defined over $\texttt{\$}$-terminated strings. Here, we define $\chi(w)$ with the $\texttt{\$}$ being implicit, not forming part of w.

Definition 5 (Measure χ [4,5]). *Let $w \in \Sigma^*$ and assume $\texttt{\$} \notin \mathcal{F}_w$. Then, $\chi(w) = |\mathcal{S}|$, where $\mathcal{S}$ is a smallest suffixient set for $w\texttt{\$}$.*

One can see from the above definitions that χ is well-defined because $\chi(w) = \texttt{sre}(w\texttt{\$})$. We will use this relation to prove results on χ via $\texttt{sre}$.

4 Sensitivity of χ to String Operations

The sensitivity to string operations has been studied for many repetitiveness measures [1,9,10,14,15,23,28,29]. It is desirable for a repetitiveness measure to not change much upon small changes in the sequence. Some repetitiveness measures are resistant to edit operations. For instance, b, z and g can only increase by a multiplicative constant after an edit operation [1], though they increase only by $O(1)$ when prepending or appending a character. On the other hand, r can increase by a $\Theta(\log n)$ factor when appending a character [15, Prop. 37]. Other results have been obtained concerning more complex string operations, like reversing a string [14], or applying a string morphism [9,10].

In this section we study how $\texttt{sre}$ and χ behave in this respect. We start by proving the following useful lemma.

Lemma 1. *If $E_r(w_1) \subseteq E_r(w_2)$, then $\texttt{sre}(w_1) \le \texttt{sre}(w_2)$.*

Proof. Let $x, y \in \mathcal{S}_r(w_1)$ with $x \neq y$. Because $x \in E_r(w_2)$, there exists $z \in \mathcal{S}_r(w_2)$ with x a suffix of z. Because y is not a suffix of x and vice versa, y cannot be a suffix of z. Therefore, the map $x \mapsto z$ with $x \in \mathcal{S}_r(w_1)$, $z \in \mathcal{S}_r(w_2)$, and $z = z'x$ for some $z' \in \Sigma^*$ is injective and then $\texttt{sre}(w_1) \le \texttt{sre}(w_2)$. □

We now prove that $\texttt{sre}(w)$ grows only by $O(1)$ when prepending or appending characters.

Lemma 2. *Let $w \in \Sigma^*$, and $c \in \Sigma$. It holds $\texttt{sre}(w) \leq \texttt{sre}(wc) \leq \texttt{sre}(w) + 2$.*

Proof. The lower bound follows from Lemma 1. For the upper bound, we analyse the new right-extensions that may arise due to appending c to w. For any fixed suffix xc of wc:

1. if xa does not appear in w for any $a \neq c$, then xc induces no new right-extensions in wc.
2. If for some $a \neq b$, xa and xb were both substrings of w, and $c \neq a$ and $c \neq b$, then xc is a new right-extension of wc induced by xc.
3. If x was always followed by $a \neq c$ in w (hence, not a right-extension), then both xa and xc are new right-extensions of wc induced by xc.

Cases 1 and 2 induce at most one new super-maximal right-extension in total for all possible xc, namely the longest right-extension in wc that is a suffix of wc. For Case 3, consider a fixed $a \in \Sigma$. For all the increasing-length suffixes $x_1c, x_2c, \ldots, x_tc$ of wc that became right-extensions together with $x_1a, x_2a, \ldots, x_ta$, one can see that the latter form a chain of suffixes of x_ta. Hence, we only have one possible new super-maximal right-extension ending with a, namely x_ta. There cannot be two of these chains for different symbols: if the suffix x is always followed by a, there cannot be a suffix y of x always followed by a different symbol b, otherwise, y is followed by a within xa, a contradiction. □

Lemma 3. *Let $w \in \Sigma^*$ and $c \in \Sigma$. It holds $\texttt{sre}(w) \leq \texttt{sre}(cw) \leq \texttt{sre}(w) + 2$.*

Proof. The lower bound follows from Lemma 1. For the upper bound, let cxa be the smallest prefix of cw that was not a right-extension of w, but is a right-extension of cw (if it exists). This means that cxa does not appear in w (otherwise, it would be a right-extension of w), so no prefix of cw of length $|cxa|$ or more is right-maximal. Hence, all prefixes of cw shorter than cxa were already right-extensions, and all prefixes longer than cxa cannot be right-extensions. Therefore, cxa together with some cxb appearing in w are the only possible new right-extensions in cw with respect to w. □

By letting $c = \$$ in Lemma 2, we relate χ to $\texttt{sre}$. This makes evident the relation between Combinatorics on words [21] with suffixient sets, via the common notion of *right-special factors* (what we call here right-maximal substrings).

Corollary 1. *Let $w \in \Sigma^*$ and $\$ \notin \mathcal{F}_w$. It holds $\texttt{sre}(w)+1 \leq \chi(w) \leq \texttt{sre}(w)+2$.*

Note that, while the value $\texttt{sre}(w)$ is non-decreasing after appending a character, this is not the case for the measure χ.

Lemma 4. *The measure χ is not monotone to appending a character.*

Proof. Consider $w = \mathtt{abaab}$. It holds $\chi(w) = 4$ and $\chi(w\mathtt{a}) = 3$. □

Now we study how much $\mathtt{sre}(w)$ can vary upon edit operations in arbitrary positions, rotations, and reversals. We will use the following famous string family.

Definition 6. *A* binary de Bruijn sequence of order $k > 0$ *[2] contains every binary string in* $\{\mathtt{a}, \mathtt{b}\}^k$ *as a substring exactly once. The length of these strings is* $n = 2^k + (k-1)$. *The set of binary de Bruijn sequences of order* k *is* $\mathtt{dB}(k)$.

Lemma 5. *It holds* $\mathtt{sre}(w) = 2^k = \Omega(n)$ *for any* $w[1\mathinner{..}n] \in \mathtt{dB}(k)$.

Proof. Let $w[1\mathinner{..}n]$ be a binary de Bruijn string of order k. By definition, w contains every binary string of length k as a substring exactly once. As all the possible pairs of strings $x\mathtt{a}$ and $x\mathtt{b}$ of length k appear in w, it follows that all the strings in $\mathcal{F}_w(k)$ are right-extensions. Moreover, each xc with $c \in \{\mathtt{a}, \mathtt{b}\}$ of length k is a super-maximal right-extension: otherwise, there would exist some $d \in \{\mathtt{a}, \mathtt{b}\}$ such that dxc and $dx\overline{c}$ are both substrings of w, which raises a contradiction since dx cannot appear twice in w. Moreover, there are no right-maximal strings of length k or greater, hence, there are no right-extensions of length greater than k. It follows that $\mathtt{sre}(w) = |\mathcal{F}_w(k)| = 2^k = \Omega(n)$. □

The following lemma uses the de Bruijn family to show that $\mathtt{sre}$ can grow by $\Omega(\log n)$ upon arbitrary edit operations and rotations.

Lemma 6. *Let* $w = \mathtt{a}^k\mathtt{b}\mathtt{a}^{k-2}\mathtt{b}x\mathtt{a}\mathtt{b}^k\mathtt{a}^{k-1} \in \mathtt{dB}(k)$ *be the lexicographically smallest binary de Bruijn sequence of order* k *[11,12]. It holds:*

1. *(Ins)* $\mathtt{sre}(w) - \mathtt{sre}(w') = 2k - 2$ *if* $w' = \mathtt{a}^{2k-2}\mathtt{b}x\mathtt{a}\mathtt{b}^k\mathtt{a}^{k-1}$,
2. *(Sub)* $\mathtt{sre}(w) - \mathtt{sre}(w') = 2k - 3$ *if* $w' = \mathtt{a}^k\mathtt{b}\mathtt{a}^{k-2}\mathtt{b}x\mathtt{a}\mathtt{b}^{k-1}\mathtt{c}\mathtt{a}^{k-1}$,
3. *(Del)* $\mathtt{sre}(w) - \mathtt{sre}(w') = 2k - 4$ *if* $w' = \mathtt{a}^k\mathtt{b}\mathtt{a}^{k-2}\mathtt{b}x\mathtt{a}\mathtt{b}^k\mathtt{c}\mathtt{a}^{k-1}$,
4. *(Rot)* $\mathtt{sre}(w) - \mathtt{sre}(w') = 2k - 2$ *if* $w' = \mathtt{b}\mathtt{a}^{k-2}\mathtt{b}x\mathtt{a}\mathtt{b}^k\mathtt{a}^{2k-1}$.

Proof. We prove each claim separately by comparing the supermaximal extensions of w' before and after performing the string operation on w' that yields w, for which $\mathtt{sre}(w) = 2^k$ by Lemma 5.

For Claim 1, note that $\mathtt{sre}(w')$ is the same as $\mathtt{sre}(\mathtt{a}^k\mathtt{b}x\mathtt{ab}^k\mathtt{a}^{k-1})$, as prepending the character $\mathtt{a}$ multiple times to this string to obtain w' never increases $\mathtt{sre}$; it only updates the supermaximal extension $\mathtt{a}^k$ to $\mathtt{a}^{k+1}$ and $\mathtt{a}^{k-1}\mathtt{b}$ to $\mathtt{a}^k\mathtt{b}$, and so on. For simplicity, we let $w' = \mathtt{a}^k\mathtt{b}x\mathtt{ab}^k\mathtt{a}^{k-1}$. The string w' does not contain substrings of length k of the form $\mathtt{a}^i\mathtt{ba}^{k-i-1}$ for $i \in [1\mathinner{..}k-2]$, nor the substring $\mathtt{ba}^{k-2}\mathtt{b}$. Note that for each of these substrings $y \in \mathcal{F}_w(k)$ with $y \notin \mathcal{F}_{w'}(k)$, the other corresponding right-extension y' in w sharing a length $k-1$ prefix with y is not a right-extension in w'. Moreover, note that all the suffixes of length $k-1$ of these y are not suffixes of one another, nor of the length $k-1$ suffixes of any of the substrings y' in w'. Hence, all $k-1$ length binary strings still appear in w' as the suffix of some length k substring that remains a right-extension in w', and hence, supermaximal extensions of w' have to be of length at least k. As each string of length k appearing in w' is unique, there are no supermaximal extensions of length greater than k. Thus, $\mathtt{sre}(w') = 2^k - 2(k-1)$ because we are losing $k-1$ pairs of supermaximal extensions of length k with respect to w. It follows that by inserting the $\mathtt{b}$ in w' to yield w, $\mathtt{sre}$ increases by $2(k-1)$.

For Claim 2, note that exactly k substrings of length k are lost when substituting the last $\mathtt{b}$ of w by $\mathtt{c}$: those of the form $\mathtt{b}^i\mathtt{a}^{k-i}$ with $i > 0$. This means that substrings ending in $\mathtt{b}^i\mathtt{a}^{k-i-1}$ with $0 < i < k$ are not right-maximal in w', hence, $2(k-1)$ supermaximal extensions are lost. Moreover, $\mathtt{b}^{k-2}\mathtt{c}$ is a new supermaximal extension, but its pair $\mathtt{b}^{k-1}$ is a suffix of $\mathtt{ab}^{k-1}$. Thus, $\mathtt{sre}(w') = 2^k - 2(k-1) + 1$.

For Claim 3, the analysis is similar to Claim 2, but in w', $\mathtt{b}^{k-1}$ remains as a supermaximal extension, so $\mathtt{sre}(w') = 2^k - 2(k-1) + 2$. For Claim 4, the analysis is similar to Claim 1, but in w', $\mathtt{ba}^{k-2}\mathtt{b}$ appears, while $\mathtt{a}^{k-1}\mathtt{b}$ does not. □

We now show that $\mathtt{sre}$ can grow by $\Omega(\sqrt{n})$ upon string reversals.

Lemma 7. *Let $k > 0$. Let $w_k = \prod_{i=0}^{k-1} \mathtt{c}\mathtt{a}^i\mathtt{b}\mathtt{a}^{k-i-1}\#_i\mathtt{a}^i\mathtt{b}\mathtt{a}^{k-i-1}\$_i$ on the alphabet $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\} \cup \bigcup_{i\in[0\,..\,k-1]}\{\#_i, \$_i\}$. It holds $\mathtt{sre}(w_k) = 5k - 1$ and $\mathtt{sre}(w_k^R) = 4k$.*

Proof. Note that, by construction, any substring of w_k containing $\#_i$ or $\$_i$ is not right-maximal. We list the supermaximal extensions of w_k:

1. $\mathtt{ba}^{k-1}$ and $\mathtt{c}$
2. $\mathtt{a}^i\mathtt{ba}^{k-i-1}\#_i$ and $\mathtt{a}^i\mathtt{ba}^{k-i-1}\$_i$ for $i \in [0\,..\,k-1]$,
3. $\mathtt{ca}^i$ and $\mathtt{ca}^{i-1}\mathtt{b}$ for $i \in [1\,..\,k-1]$,
4. $\mathtt{a}^i\mathtt{ba}^{k-i-1}$ for $i \in [1\,..\,k-1]$.

This sums to a total of $5k-1$ supermaximal extensions in w_k. In the reversed string $w_k^R = \prod_{i=0}^{k-1} \$_{k-i-1}\mathtt{a}^i\mathtt{ba}^{k-i-1}\#_{k-i-1}\mathtt{a}^i\mathtt{ba}^{k-i-1}\mathtt{c}$, we have instead:

1. $\mathtt{ba}^{k-1}$, $\$_{k-1}$ and $\mathtt{a}^{k-2}\mathtt{c}\$_{k-2}$
2. $\mathtt{a}^i\mathtt{ba}^{k-i-1}\#_{k-i-1}$ and $\mathtt{a}^i\mathtt{ba}^{k-i-1}\mathtt{c}$ for $i \in [0\,..\,k-1]$,
3. $\mathtt{a}^{k-i-1}\mathtt{c}\$_{k-i-2}$ for $i \in [1\,..\,k-2]$,
4. $\mathtt{a}^i\mathtt{ba}^{k-i-1}$ for $i \in [1\,..\,k-1]$.

This sums to a total of $4k$ supermaximal extensions in w_k^R, of length $|w_k| = |w_k^R| = k(2k+3)$. Thus, $\mathtt{sre}(w_k) - \mathtt{sre}(w_k^R) = k - 1 = \Theta(\sqrt{n})$. □

Finally, we show upper bounds on the sensitivity of $\mathtt{sre}$ to string operations.[5]

Lemma 8. *Let $w \in \Sigma^*$ and $w' \in \mathtt{ins}_\Sigma(w) \cup \mathtt{del}_\Sigma(w) \cup \mathtt{sub}_\Sigma(w) \cup \mathcal{R}(w) \cup \{w^R\}$. It holds*

$$\mathtt{sre}(w') - \mathtt{sre}(w) = O\left(\delta \max\left(1, \log(n/\delta\log\delta)\right)\log\delta\right) \text{ and}$$
$$\mathtt{sre}(w') \,/\, \mathtt{sre}(w) = O\left(\max\left(1, \log(n/\delta\log\delta)\right)\log\delta\right).$$

Proof. It follows because the multiplicative sensitivity of δ to the string operations considered and reversals is $O(1)$ [1], the known relations $\delta \le \gamma \le \chi \le 2\overline{r}$ [4], and the upper bound $r = O(\delta\max(1, \log(n/\delta\log\delta))\log\delta)$ [17]. □

[5] For the multiplicative sensitivity, we assume w and w' are not unary strings, as otherwise $\mathtt{sre}(w)$ or $\mathtt{sre}(w')$ would be 0. This does not happen with χ.

5 Relating χ to Other Repetitiveness Measures

Previous work [4] established that $\gamma = O(\chi)$ and $\chi = O(\bar{r})$ on every string family. In this section we obtain the more natural result that χ is always $O(r)$, and that it can be asymptotically strictly smaller, $\chi = o(r)$, on some string families (we actually prove $\chi = o(v)$). We also show that χ is incomparable with all the copy-paste measures except b, in the sense that there are string families where χ is asymptotically strictly smaller than each other, and vice versa.

5.1 Proving $\chi = O(r)$

We first prove that χ is asymptotically upper-bounded by the number r of runs in the BWT of the sequence.

Lemma 9. *It always holds that $\chi \leq 2r$.*

Proof. Let xa be a super-maximal right-extension in $w\$$. We distinguish, in the BWT-matrix of $w\$$, the sets $\mathcal{R}_{xc}$ of rotations starting with xc where $c \in \Sigma \cup \{\$\}$. Because x is right-maximal, at least 2 of these blocks are non-empty; i) the set $\mathcal{R}_{xa}$; ii) some set $\mathcal{R}_{xb}$ of rotations starting with xb, where $b \neq a$.

We claim that the last characters of the rotations in $\mathcal{R}_{xa}$ must be disjoint from the last characters of rotations in $\mathcal{R}_{xb}$, for any $b \neq a$. Suppose by contradiction that there are two rotations of $w\$$, of the form $xa \ldots c$ and $xb \ldots c$. Then, cxa and cxb are circular substrings of $w\$$. Note that cx does not contain $\$$, otherwise, as $\$$ is unique and both circular substrings have the same length, cxa and cxb would have to be the same string, yet a and b are different. This implies cxa and cxb have to be substrings of $w\$$. Moreover, cx is a right-maximal substring of $w\$$, and cxa is a one-character right-extension that contains xa as a suffix, contradicting that xa is super-maximal.

Let $a_1 < \cdots < a_t$ be all the characters such that xa_i is a super-maximal right-extension, for $t \geq 2$. Those induce (not necessarily consecutive) BWT areas $\mathcal{R}_{xa_i}$. The argument of the previous paragraph applies to any pair $a = a_i$ and $b = a_{i+1}$, and implies that a new BWT run must start between the first row following $\mathcal{R}_{xa_i}$ and the first row of $\mathcal{R}_{xa_{i+1}}$, for all $i = 2, \ldots, t$. The string x then induces t super-maximal right-extensions and $t-1$ BWT runs. The worst ratio between contributions to `sre` and to r is 2 to 1, which occurs when $t = 2$.

The contributions can be summed for other super-maximal right-extensions ya'_i if y does not prefix or is prefixed by x, because the corresponding BWT ranges are disjoint. We now focus in the case where y is a prefix of x. Since y is right-maximal, its range $\mathcal{R}_y$ strictly contains $\mathcal{R}_x$. Further, $\mathcal{R}_x$ is completely contained in the range of one of the extensions of y, precisely $\mathcal{R}_{ye}$ where $e = x[|y|+1]$. The characters $a'_1 < \cdots < a'_{t'}$ such that ya'_i is a super-maximal right-extension also induce $t'-1$ BWT runs, starting between the row following the areas $\mathcal{R}_{ya'_i}$ and the first row of the areas $\mathcal{R}_{ya'_{i+1}}$. Importantly, those induced run start positions are disjoint from those induced by x: the first run start position induced by x is after $\mathcal{R}_{xa_1}$, whereas the only run start induced by y inside $\mathcal{R}_{ye}$ is at the first row of $\mathcal{R}_{xa_1}$ or earlier. Therefore, the induced runs of x and y are disjoint too.

Adding over all super-maximal extensions, it follows that $r \geq \chi/2$. □

5.2 A family with $\chi = o(v)$ (and thus $o(r)$)

We will now show that $\chi = o(v)$ on the so-called Fibonacci words, which also implies $\chi = o(r)$ in that string family because $v = O(r)$ [27]. Combined with Lemma 9, this implies that χ is a strictly smaller measure than r. Instead, χ is incomparable with v, as we show later. In our way, we obtain some relevant byproducts about the structure of suffixient sets on Fibonacci, and more generally, episturmian words.

Definition 7 ([8,16]). *An infinite string $\boldsymbol{w}$ is* episturmian *if it has at most one right-maximal substring of each length and its set of substrings is closed under reversal, that is, $\mathcal{F}_{\boldsymbol{w}} = \mathcal{F}_{\boldsymbol{w}}^R$. It is* standard episturmian *(or* epistandard*) if, in addition, all the right-maximal substrings of $\boldsymbol{w}$ are of the form $\boldsymbol{w}[1\mathinner{..}i]^R$ with $i \geq 0$, i.e., they are the reverse of some prefix of $\boldsymbol{w}$.*

Lemma 10. *Let $\boldsymbol{w} \in \Sigma^\omega$ be an episturmian word with $\sigma \geq 2$. Then, $\mathtt{sre}(\boldsymbol{w}[i\mathinner{..}j]) \leq \sigma$ for $i, j \geq 0$.*

Proof. Let $\mathbf{w}$ be an epistandard word. The right-extensions $x_1, x_2, \dots$ ending with $a \in \Sigma$ form a *suffix-chain* where each x_i is a suffix of x_{i+1}. There is one of those suffix-chains for each character $a \in \Sigma$.

Let $\mathbf{w}$ be episturmian but not epistandard. There exists some epistandard word $\mathbf{s}$ with the same set of substrings, i.e., $\mathcal{F}_{\mathbf{w}} = \mathcal{F}_{\mathbf{s}}$ [8]. Therefore, for any episturmian word $\mathbf{w}$, there exist exactly σ suffix-chains of right-extensions.

It follows that for any substring $\mathbf{w}[i\mathinner{..}j]$ of any episturmian word $\mathbf{w}$, $\mathtt{sre} \leq \sigma$, and the supermaximal extension for each $a \in \Sigma$ appearing in $\mathbf{w}[i\mathinner{..}j]$ is the longest reversed prefix of $\mathbf{w}$ followed by a appearing in $\mathbf{w}[i\mathinner{..}j]$, and having another occurrence within $\mathbf{w}[i\mathinner{..}j]$ followed by another character. □

Combining this result with Lemma 2, we have the following bound.

Corollary 2. *For any episturmian word $\boldsymbol{w} \in \Sigma^\omega$ it holds $\chi(\boldsymbol{w}[i\mathinner{..}j]) \leq \sigma + 2$.*

The next lemma precisely characterizes the suffixient sets of Fibonacci words, a particular case of epistandard words that will be useful to relate χ with v.

Definition 8. *Let $F_1 = \mathtt{b}$, $F_2 = \mathtt{a}$, and $F_k = F_{k-1}F_{k-2}$ for $k \geq 3$ be the Fibonacci family of strings. Their lengths, $f_k = |F_k|$, form the Fibonacci sequence.*

Lemma 11. *Every Fibonacci word $F_k\$$ has a suffixient set of size at most 4. For $k \geq 6$, the only smallest suffixient sets for $F_k\$$ are $\{f_k+1, f_k-1, f_{k-1}-1, p\}$, where $p \in \{f_{k-2}+1, 2f_{k-2}+1\}$.*

Proof. The upper bound of 4 stems directly from Corollary 2, because the infinite Fibonacci word is binary epistandard. For $k \geq 3$, there exist strings H_k such that $F_k = F_{k-1}F_{k-2} = H_k cd$ and $F_{k-2}F_{k-1} = H_k dc$, for $cd = \mathtt{ab}$ or $cd = \mathtt{ba}$ depending on the parity of k [22]. Let us call $F'_k = H_k dc = F_{k-2}F_{k-1}$, that is, F_k with the last two letters exchanged; thus $F_k = F_{k-1}F_{k-2} = F_{k-2}F'_{k-1}$.

Note that $F_{k-1} = H_{k-1}dc$ prefixes F_k. On the other hand, we can write $F_k = F_{k-1}F_{k-2} = F_{k-2}F_{k-3}F_{k-2} = F_{k-2}F'_{k-1} = F_{k-2}H_{k-1}cd$. Therefore, string H_{k-1} is right-maximal in F_k. Its extensions, $H_{k-1}d$ and $H_{k-1}c$, are super-maximal because there are no other occurrences of H_{k-1} in F_k: (i) H_{k-1} cannot occur starting at positions $f_{k-2}+2$ or $f_{k-2}+3$ because it occurs at $f_{k-2}+1$, so H_{k-1} should match itself with an offset of 1 or 2, which is impossible because it prefixes F_{k-1} and all F_{k-1} for $k-1 \geq 5$ start with `abaab`; (ii) H_{k-1} cannot occur starting at positions 2 to f_{k-2} because its prefix F_{k-2} should occur inside the prefix $F_{k-2}F_{k-2}$ of $F_k = F_{k-2}F'_{k-1} = F_{k-2}F_{k-2}F'_{k-3}$, and so F_{k-2} should equal a rotation of it, which is impossible [7, Cor. 3.2]. The two positions following H_{k-1}, $f_{k-1}-1$ and $f_k - 1$, then appear in any suffixient set.

On the other hand, F_{k-2} is followed by `$` in F_k`$`, and it also prefixes $F_k = F_{k-2}F'_{k-1}$, therefore F_{k-2} is right-maximal. The first occurrence is preceded by F_{k-1}, and hence by c, and the second by no symbol. F_{k-2} also occurs in F_k at position $f_{k-2}+1$, as seen above, preceded by F_{k-2} and thus by d. There are no other occurrences of F_{k-2} in F_k because (i) it cannot occur starting at positions 2 to f_{k-2} by the same reason as point (ii) of the previous paragraph; (ii) it cannot appear starting at positions $f_{k-2}+2$ to $f_{k-1}-2$ because $F_k = F_{k-2}F_{k-2}F'_{k-3}$ and $F'_{k-3}[1, f_{k-3}-2] = F_{k-3}[1, f_{k-3}-2] = F_{k-2}[1..f_{k-3}-2]$, thus such an occurrence would also match a rotation of F_{k-2}, which is impossible as noted above; (iii) it cannot appear starting at positions $f_{k-1}-1$ or f_{k-1} because, since it matches at position $f_{k-1}+1$, F_{k-2} would match itself with an offset of 1 or 2, which is impossible as noted in point (i) of the previous paragraph. The right-extensions of F_{k-2} are then super-maximal. The one followed by `$` occurs ending at position $f_k + 1$. The other two are followed by `a` because they are followed by F_{k-2} and by F'_{k-3} and all F_k for $k \geq 2$ start with `a`. We can then choose either ending position for a suffixient set, $f_{k-2}+1$ or $2f_{k-2}+1$. □

Corollary 3. *There exist string families where $\chi = o(v)$.*

Proof. It follows from Lemma 11 and the fact that $v = \Omega(\log n)$ on the odd Fibonacci words [27, Thm. 28].

5.3 Uncomparability of χ with copy-paste measures

Finally, we show that χ is incomparable with most copy-paste measures. This follow from χ being $\Theta(n)$ on de Bruijn sequences and $O(1)$ on Fibonacci strings. Because $g = O(n/\log n)$ on de Bruijn sequences [27] and by Lemma 5, we have:

Corollary 4. *There exists a string family with $\chi = \Omega(g \log n)$.*

This result is particularly relevant because all the copy-paste based measures μ, with the exception of z_e, are $O(g)$. Corollary 4 then implies $\mu = o(\chi)$ on de Bruijn sequences for all these measures μ.

While it has been said that $z_e = O(n/\log n)$ on binary sequences as well [19], this referred to the version that adds to each phrase the next nonmatching character. Because z_e is not an optimal parse, it is not obvious that this also holds for the version studied later in the literature, which does not add the next character. We then prove next that $z_e = o(\chi)$ holds on de Bruijn words.

Lemma 12. *There exists a string family with* $\chi = \Omega\left(z_e \frac{\log n \log\log\log n}{(\log\log n)^2}\right)$.

Proof. It always holds that $z_e = O\left(z \frac{\log^2(n/z)}{\log\log(n/z)}\right)$ [13]. In de Bruijn sequences it holds that $z = \Theta(n/\log n)$, so $n/z = \Theta(\log n)$. Therefore, $z_e = O\left(z \frac{(\log\log n)^2}{\log\log\log n}\right)$, and replacing $z = \Theta(n/\log n)$ we get $z_e = O\left(n \frac{(\log\log n)^2}{\log n \log\log\log n}\right)$. By Lemma 5, this yields $\chi = \Omega\left(z_e \frac{\log n \log\log\log n}{(\log\log n)^2}\right) = \omega(z_e)$ on de Bruijn sequences. □

Corollary 5. *The measure* χ *is uncomparable to* $\mu \in \{z, z_{no}, z_e, z_{end}, v, g, g_{rl}, c\}$.

Proof. From Corollary 4 and Lemma 12, and that z, z_{no}, z_{end}, v, g_{rl} and c are always $O(g)$, it follows that there are string families where $\mu = o(\chi)$, for any $\mu \in \{z, z_{no}, z_e, z_{end}, v, g, g_{rl}, c\}$. On the other hand, from Lemma 11 and Corollary 3, and that $c = \Omega(\log n)$ on Fibonacci words [27, Thm. 32] and $c = O(\mu)$ for any $\mu \in \{z, z_{no}, z_e, z_{end}, g_{rl}, g\}$ [27, Thm. 30], it follows that there are string families where $\chi = o(\mu)$, for any $\mu \in \{z, z_{no}, z_e, z_{end}, v, g, g_{rl}, c\}$. □

6 Conclusions and Open Questions

We have contributed to the understanding of χ as a new measure of repetitiveness, better finding its place among more studied ones. Figure 1 shows the (now) known relations around χ (cf. [26]).

There are still many interesting open questions about χ. One of the most important is whether χ is reachable. Proving $b = O(\chi)$ would settle this question on the affirmative, and at the same time give the first copy-paste measure that is comparable with χ. We conjecture, instead, that χ is not reachable, proving which would imply that γ is also unreachable, a long-time open question.

One consequence of Corollary 4 is that $\chi \notin O(g \log^k(n/g))$ for any $k > 0$. It could be the case, though, that $\chi = O(\delta \log n)$, because the separation of χ and δ on de Bruijn sequences is a $\Theta(\log n)$ factor.

Regarding edit operations, it seems that that $\texttt{sre}(w')/\texttt{sre}(w)$ is $O(1)$ for all the string operations we considered. Showing a multiplicative constant for insertion would imply the existence of a constant for rotation and vice versa. It is also open whether $r = O(\chi \log \chi)$. If this were true —and provided that χ has $O(1)$ multiplicative sensitivity to string operations— it would imply that r has $O(\log n)$ multiplicative sensitivity to these operations, making the already known lower bounds on multiplicative sensitivity [1,14,15] tight. If the conjecture were false, then χ could be considerably smaller than r in some string families.

References

1. Akagi, T., Funakoshi, M., Inenaga, S.: Sensitivity of string compressors and repetitiveness measures. Information and Computation **291**, 104999 (2023)
2. Bruijn, de, N.: A combinatorial problem. Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam **49**(7), 758–764 (1946)
3. Burrows, M., Wheeler, D.: A block sorting lossless data compression algorithm. Tech. Rep. 124, Digital Equipment Corporation (1994)
4. Cenzato, D., Depuydt, L., Gagie, T., Kim, S.H., Manzini, G., Olivares, F., Prezza, N.: Suffixient arrays: a new efficient suffix array compression technique. CoRR **2407.18753** (2025)
5. Cenzato, D., Olivares, F., Prezza, N.: On computing the smallest suffixient set. In: Proc. 31st International Symposium on String Processing and Information Retrieval (SPIRE 2024). Lecture Notes in Computer Science, vol. 14899, pp. 73–87. Springer (2024)
6. Depuydt, L., Gagie, T., Langmead, B., Manzini, G., Prezza, N.: Suffixient sets. CoRR **2312.01359** (2023)
7. Droubay, X.: Palindromes in the Fibonacci word. Information Processing Letters **55**(4), 217–221 (1995)
8. Droubay, X., Justin, J., Pirillo, G.: Episturmian words and some constructions of de Luca and Rauzy. Theoretical Computer Science **255**(1), 539–553 (2001)
9. Fici, G., Romana, G., Sciortino, M., Urbina, C.: On the impact of morphisms on BWT-runs. In: Proc. 34th Annual Symposium on Combinatorial Pattern Matching (CPM 2023). LIPIcs, vol. 259, pp. 10:1–10:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2023)
10. Fici, G., Romana, G., Sciortino, M., Urbina, C.: Morphisms and BWT-run sensitivity. CoRR **2504.17443** (2025)
11. Fredricksen, H.: A survey of full length nonlinear shift register cycle algorithms. SIAM Review **24**(2), 195–221 (1982)
12. Gabric, D., Sawada, J., Williams, A., Wong, D.: A framework for constructing de Bruijn sequences via simple successor rules. Discrete Mathematics **341**(11), 2977–2987 (2018)
13. Gawrychowski, P., Kosche, M., Manea, F.: On the number of factors in the LZ-end factorization. In: Proc. 30th International Symposium on String Processing and Information Retrieval (SPIRE 2023). Lecture Notes in Computer Science, vol. 14240, pp. 253–259. Springer (2023)
14. Giuliani, S., Inenaga, S., Lipták, Z., Prezza, N., Sciortino, M., Toffanello, A.: Novel results on the number of runs of the Burrows-Wheeler-Transform. In: Proc. 47th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2021). Lecture Notes in Computer Science, vol. 12607, pp. 249–262. Springer (2021)
15. Giuliani, S., Inenaga, S., Lipták, Z., Romana, G., Sciortino, M., Urbina, C.: Bit catastrophes for the Burrows-Wheeler transform. Theory of Computing Systems **69**(2), 19 (2025)
16. Glen, A., Justin, J.: Episturmian words: a survey. RAIRO - Theoretical Informatics and Applications **43**(3), 403–442 (2009)
17. Kempa, D., Kociumaka, T.: Resolution of the Burrows-Wheeler transform conjecture. Communications of the ACM **65**(6), 91–98 (2022)

18. Kempa, D., Prezza, N.: At the roots of dictionary compression: String attractors. In: Proc. 50th Annual ACM Symposium on the Theory of Computing (STOC 2018). pp. 827–840. ACM (2018)
19. Kreft, S., Navarro, G.: On compressing and indexing repetitive sequences. Theoretical Computer Science **483**, 115–133 (2013)
20. Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE Transactions on Information Theory **22**(1), 75–81 (1976)
21. Lothaire, M.: Algebraic Combinatorics on Words. Encyclopedia of Mathematics and its Applications, Cambridge University Press, New York, NY, USA (2002)
22. de Luca, A.: A combinatorial property of the Fibonacci words. Information Processing Letters **12**(4), 193–195 (1981)
23. Mantaci, S., Restivo, A., Romana, G., Rosone, G., Sciortino, M.: A combinatorial view on string attractors. Theoretical Computer Science **850**, 236–248 (2021)
24. Navarro, G.: Indexing highly repetitive string collections, part I: Repetitiveness measures. ACM Computing Surveys **54**(2), article 29 (2021)
25. Navarro, G.: Indexing highly repetitive string collections, part II: Compressed indexes. ACM Computing Surveys **54**(2), article 26 (2021)
26. Navarro, G.: Indexing highly repetitive string collections. CoRR **2004.02781** (2022)
27. Navarro, G., Ochoa, C., Prezza, N.: On the approximation ratio of ordered parsings. IEEE Transactions on Information Theory **67**(2), 1008–1026 (2021)
28. Navarro, G., Olivares, F., Urbina, C.: Generalized straight-line programs. Acta Informatica **62**(1), 14 (2025)
29. Navarro, G., Urbina, C.: Repetitiveness measures based on string morphisms. Theoretical Computer Science **1043**, 115259 (2025)
30. Storer, J.A., Szymanski, T.G.: Data compression via textual substitution. Journal of the ACM **29**(4), 928–951 (1982)