Theory of Computing Systems (2025) 69:19
https://doi.org/10.1007/s00224-024-10212-9

Bit Catastrophes for the Burrows-Wheeler Transform

Sara Giuliani[1] · Shunsuke Inenaga[2] · Zsuzsanna Lipták[1] · Giuseppe Romana[3] · Marinella Sciortino[3] · Cristian Urbina[4]

Accepted: 29 December 2024

Abstract

A bit catastrophe, loosely defined, is when a change in just one character of a string causes a significant change in the size of the compressed string. We study this phenomenon for the Burrows-Wheeler Transform (BWT), a string transform at the heart of several of the most popular compressors and aligners today. The parameter determining the size of the compressed data is the number of equal-letter runs of the BWT, commonly denoted r. We exhibit infinite families of strings in which insertion, deletion, resp. substitution of one character increases r from constant to $\Theta(\log n)$, where n is the length of the string. These strings can be interpreted both as examples for an increase by a multiplicative or an additive $\Theta(\log n)$-factor. As regards the multiplicative factor, they attain the upper bound given by Akagi, Funakoshi, and Inenaga [Inf & Comput. 2023] of $\mathcal{O}(\log n \log r)$, since here $r = \mathcal{O}(1)$. We then give examples of strings in which insertion, deletion, resp. substitution of a character increases r by a $\Theta(\sqrt{n})$ additive factor. These strings significantly improve the best known lower bound for an additive factor of $\Omega(\log n)$ [Giuliani et al., SOFSEM 2021].

✉ Marinella Sciortino
marinella.sciortino@unipa.it

Sara Giuliani
sara.giuliani_01@univr.it

Shunsuke Inenaga
inenaga.shunsuke.380@m.kyushu-u.ac.jp

Zsuzsanna Lipták
zsuzsanna.liptak@univr.it

Giuseppe Romana
giuseppe.romana01@unipa.it

Cristian Urbina
crurbina@dcc.uchile.cl

1 University of Verona, Verona, Italy

2 Kyushu University, Fukuoka, Japan

3 University of Palermo, Palermo, Italy

4 University of Chile, Santiago, Chile

Published online: 15 April 2025

Keywords Burrows-Wheeler transform · Equal-letter run · Repetitiveness measure · Sensitivity

1 Introduction

The Burrows-Wheeler Transform (BWT) [9] is a reversible transformation of a string, consisting of a permutation of its characters. It can be obtained by sorting all of its rotations and then concatenating their last characters. The BWT is present in several compressors, such as *bzip* [38]. It also lies at the heart of some of the most powerful compressed indexes in terms of query time and functionality, such as the well-known *FM-index* [15], and the more recent *RLFM-index* [30, 31], *r-index* [3, 17, 18], and *extended r-index* [7], in which extensions of the BWT to string collections are also used [6, 12, 32]. Some of the most commonly used bioinformatics tools such as `bwa` [28, 39], `bowtie` [26, 27], and SOAP2 [25] also use the BWT at their core.

Given a string w, the measure $r = r(w)$ is defined as $r(w) = runs(\mathrm{BWT}(w))$, where $runs(v)$ denotes the number of maximal equal-letter runs of a string v. It is well known that r tends to be small on repetitive inputs. This is because, on texts with many repeated substrings, the BWT tends to create long runs of equal characters (so-called *clustering effect*) [36]. Due to the widespread use of runlength-compressed BWT-based data structures, the BWT can thus be viewed as a *compressor*, with the number of runs r the compression size. On the other hand, r is also increasingly being used as a *repetitiveness measure*, i.e. as a parameter of the input text itself. In his recent survey [35], Navarro explored the relationships between many repetitiveness measures, including r. In particular, all repetitiveness measures considered are lower bounded by δ [23], a measure closely related to the *factor complexity* function [29]. It was further shown in [21] that r is within a polylog(n) factor of z, the number of phrases of the LZ77-compressor [40].

Giuliani et al. [19] studied the ratio of the runs of the BWT of a string and of its reverse. The authors gave an infinite family of strings for which this ratio is $\Theta(\log n)$, where n is the length of the string. This family can also serve as an example of strings in which appending one character can cause r to increase from $\Theta(1)$ to $\Theta(\log n)$. In this paper, we further explore this effect, extending it to the other edit operations of deletion and substitution, for which we also give examples of a change of r from $\Theta(1)$ to $\Theta(\log n)$. This attains the known upper bound of $\mathcal{O}(\log r \log n)$ [1]. Besides proving the bound's tightness, we also explicitly compute the output of the BWT for this family of strings, after a single edit operation has been applied. This is of potential independent interest as it can be connected to the sensitivity of other repetitiveness measures closely related to r, such as the r_B measure, defined as the number of runs of the bijective BWT [2, 5].

Akagi et al. [1] explored the question of how changes of just one character affect the compression ratio of known compressors; they refer to this as the compressors' *sensitivity*. More precisely, the sensitivity of a compressor is the maximum *multiplicative factor* by which a single-character edit operation can change the size of the output of the compressor. In addition, they also study the maximum *additive factor* an edit operation may cause in the output. Our second family of strings falls in this category:

these are strings with r in $\Theta(\sqrt{n})$ on which an edit operation (insertion, deletion, or substitution) can cause r to increase by a further additive factor of $\Theta(\sqrt{n})$. This is a significant improvement over the previous lower bound of $\Omega(\log n)$ [19]. We explicitly compute how the BWT changes after a single edit operation has been applied. This is again of potential independent interest; possible applications include studying the tightness of the known upper bound $\mathcal{O}(r \log r \log n)$ [1], or comparing the sensitivity of other measures with respect to the same family of strings.

Lagarde and Perifel in [24] show that Lempel-Ziv 78 (LZ78) compression [41] suffers from the so-called "one-bit catastrophe": they give an infinite family of strings for which prepending a character causes a previously compressible string to become incompressible. They also show that this "catastrophe" is not a "tragedy": they prove that this change from compressible to incompressible can only happen when the original string was already poorly compressible. They also present a family of strings which are maximally compressible but exhibit a strong case of bit catastrophe, while still remaining compressible.

Here we use the term "one-bit catastrophe" in a looser meaning, namely simply to denote the effect that an edit operation may change the compression size significantly, i.e. increase it such that $r(w'_n) = \omega(r(w_n))$[1] , for an infinite family $(w_n)_{n>0}$, where w'_n is the word resulting from applying a single edit operation to w_n. For a stricter terminology we would need to decide for one of the different definitions of BWT-compressibility currently in use. In particular, a string may be called compressible with the BWT if r is in $\mathcal{O}(n/\text{polylog}(n))$ [21], or if $r(w)/runs(w) \to 0$ [16], or even as soon as $runs(w) > r(w)$ [33].

Note that, in contrast to Lempel-Ziv compression, for the BWT, appending, prepending, and inserting are equivalent operations, since the BWT is invariant w.r.t. conjugacy. This means that, if there exists a word w and a character c s.t. appending c to w causes a certain change in r, then this immediately implies the existence of equivalent examples for prepending and inserting character c. This is because $r(wc)/r(w) = x$ (appending) implies that $r(cw)/r(w) = x$ (prepending), as well as $r(ucv)/r(uv) = x$, for every conjugate uv of $w = vu$ (insertion).

Finally, the BWT comes in two variants: in one, the transform is applied directly on the input string: this is the preferred variant in the literature on combinatorics on words, and the one we concentrate on in most of the paper. In the other, the input string is assumed to have an end-of-string marker, usually denoted \$: this variant is common in the string algorithms and data structures literature. We show that there can be a multiplicative $\Theta(\log n)$, or an additive $\Theta(\sqrt{n})$ factor difference between the two transforms. It is interesting to note that the previous remark about the equivalence of insertion in different places in the text does not extend to the variant with the final dollar. We show, however, that our results regarding the $\Theta(\sqrt{n})$ additive factor apply also to this variant, for all three edit operations, and that appending a character at the end of the string—i.e., just before the \$-character—can result in a multiplicative $\Theta(\log n)$ increase. This is in stark contrast with the known fact that *prepending* a character can change the number of runs of the \$-variant by at most 2 [1].

[1] We say that $f(n) = \omega(g(n))$ if $\lim_{n\to\infty} f(n)/g(n) = \infty$.

This work is an extended version of our conference article with the same title, published in the proceedings of DLT 2023 [20]. Specifically, we included all proofs (most of which had to be omitted in the conference version due to space limitations) and restructured parts of the paper, including several new results.[2]

2 Preliminaries

In this section, we give the necessary definitions and terminology used throughout the paper.

2.1 Basics on Words

Let Σ be a finite ordered *alphabet*, of cardinality σ. The elements of Σ are called *characters* or *letters*. A *word* (or *string*) over Σ is a finite sequence $w = w[0]w[1]\cdots w[n-1] = w[0..n-1]$ of characters from Σ. We denote by $|w| = n$ the length of w, with ϵ the unique word of length 0. The set of words of length n is denoted Σ^n, and $\Sigma^* = \cup_{n\geq 0}\Sigma^n$ is the set of all words over Σ. Given a word $w = w[0..n-1]$, its *reverse* is the word $\mathrm{rev}(s) = w[n-1]w[n-2]\cdots w[0]$. For a non-empty word $w = w[0..n-1]$, we denote by $\widehat{w}$ the word $w[0..n-2]$, i.e. w without its last character. We use the notation $\prod_{i=1}^{k} v_i$ for the concatenation of the words $v_1, v_2, \ldots, v_k$. In particular, v^k for $k \geq 1$ stands for the k-fold concatenation of the word v.

Let w be a word over Σ. If $w = uxv$ for some (possibly empty) words $u, x, v \in \Sigma^*$, then u is called a *prefix*, x a *substring* (or *factor*), and v a *suffix* of w. A *proper* prefix, substring, or suffix of w is one that does not equal w. If x is a substring of w, then there exist i, j such that $x = w[i]\cdots w[j] = w[i..j]$, where $w[i..j] = \epsilon$ if $i > j$. If $w[i..j] = x$, then i is called an *occurrence* of x.

Let $u, v \in \Sigma^*$. If $w = uv$ and $w' = vu$, then w and w' are called *conjugates* or *rotations* of one another. Equivalently, w' is a conjugate of w if there is $0 \leq i \leq |w|$ such that $w' = w[i..|w|]w[0..i-1]$. In this case, we write $w' = \mathrm{conj}_i(w)$. A word u is a *circular factor* of a word w if it is a prefix of $conj_i(w)$ for some $0 \leq i < |w|$, and i is called an *occurrence* of u. If a word w can be written as $w = v^k$ for some $k > 1$, then w is called a *power*, otherwise w is called *primitive*. It is easy to see that w is primitive if and only if it has $|w|$ distinct conjugates.

Given two words v, w, the *longest common prefix* of v and w, denoted $\mathrm{lcp}(v, w)$, is the unique word u such that u is a prefix of both v and w, and $v[|u|] \neq w[|u|]$ if neither of the two words is prefix of the other. The *lexicographic order* between two distinct words $v, w \in \Sigma^*$ is defined as follows: $v <_{\mathrm{lex}} w$ (or simply $v < w$) if $v = \mathrm{lcp}(v, w)$, or else if $v[|u|] < w[|u|]$, where $u = \mathrm{lcp}(v, w)$. A word is called a *Lyndon word* if it is lexicographically strictly smaller than all of its conjugates.

[2] Section 3 has been extended by an exploration of some structural properties of Fibonacci words and central words (Lemmas 5, 9, 10 and Cor. 11), which could be of independent interest. We also assess how the BWT changes when appending or prepending a character to a Lyndon word (Lemma 16), and we now include the explicit form of the BWT for all three types of bit catastrophe on Fibonacci words, and not only the number of runs (Props. 3, 4, 17). Finally, we have added the multiplicative bit catastrophe for $r_{\$}$ in Prop. 37.

Finally, an *equal-letter run* (or simply *run*) is a maximal substring consisting of the same character, and $runs(v)$ is the number of equal-letter runs in the word v. For example, the word `catastrophic` has 12 runs, while the word `mississippi` has 8 runs.

2.2 The Burrows-Wheeler Transform

Let $w \in \Sigma^*$. The *conjugate array* $\mathrm{CA} = \mathrm{CA}_w$ of w is a permutation of $\{0, 1, \ldots, n-1\}$ defined by: $\mathrm{CA}[i] < \mathrm{CA}[j]$ if (i) $\mathrm{conj}_i(w) <_{\mathrm{lex}} \mathrm{conj}_j(w)$, or (ii) $\mathrm{conj}_i(w) = \mathrm{conj}_j(w)$ and $i < j$. So $\mathrm{CA}[k]$ contains the index of the kth conjugate of w in lexicographic order. (Note that the conjugate array is the circular equivalent of the suffix array.) For example, if $w =$ `catastrophic`, then $\mathrm{CA}_w = [3, 1, 0, 11, 9, 10, 7, 8, 6, 4, 2, 5]$.

The *Burrows-Wheeler Transform* (BWT) of the word w is a permutation of the characters of w, usually denoted $L = \mathrm{BWT}(w)$, defined as $L[i] = w[n-1]$ if $\mathrm{CA}[i] = 0$, and $L[i] = w[\mathrm{CA}[i] - 1]$ otherwise. It follows from the definition that w and w' are conjugates if and only if $\mathrm{BWT}(w) = \mathrm{BWT}(w')$. In Fig. 1a we exhibit the list of all lexicographically sorted rotations of the word $w =$ `catastrophic` (referred to as *BWT matrix*), as well as the conjugate array of w. The BWT of w is the last column of this matrix, namely `tcciphrotaas`.

We denote with $r(w) = runs(\mathrm{BWT}(w))$ the number of runs in the BWT of the word w. For example, $r($`catastrophic`$) = runs($`tcciphrotaas`$) = 10$.

In the context of string algorithms and data structures, it is usually assumed that each string terminates with an end-of-string symbol (denoted by \$), not occurring elsewhere in the string; the \$-symbol is smaller than all other symbols in the alphabet. In fact, with this assumption, sorting the conjugates of w\$ can be reduced to lexicographically sort-

	CA	rotations of `catastrophic`	L
0	3	`astrophiccat`	`t`
1	1	`atastrophicc`	`c`
2	0	`catastrophic`	`c`
3	11	`ccatastrophi`	`i`
4	9	`hiccatastrop`	`p`
5	10	`iccatastroph`	`h`
6	7	`ophiccatastr`	`r`
7	8	`phiccatastro`	`o`
8	6	`rophiccatast`	`t`
9	4	`strophiccata`	`a`
10	2	`tastrophicca`	`a`
11	5	`trophiccatas`	`s`

(a)

	CA	rotations of `catastrophic$`	L
0	12	`$catastrophic`	`c`
1	3	`astrophic$cat`	`t`
2	1	`atastrophic$c`	`c`
3	11	`c$catastrophi`	`i`
4	0	`catastrophic$`	`$`
5	9	`hic$catastrop`	`p`
6	10	`ic$catastroph`	`h`
7	7	`ophic$catastr`	`r`
8	8	`phic$catastro`	`o`
9	6	`rophic$catast`	`t`
10	4	`strophic$cata`	`a`
11	2	`tastrophic$ca`	`a`
12	5	`trophic$catas`	`s`

(b)

Fig. 1 The BWT matrix of the word $w =$`catastrophic` is shown in the table on the left (a). The second column, denoted by CA, gives the conjugate array of w, the last column shows the BWT of w. The BWT matrix of the word `catastrophic$` (the character `$` is appended to w) is shown on the right (b)

ing its suffixes. Note that appending the character \$ to the word w changes the output of BWT. We denote by $r_{\$}(w) = runs(\text{BWT}(w\$))$. For example, as shown in Fig. 1b, $\text{BWT}(\texttt{catastrophic\$}) = \texttt{ctci\$phrotaas}$ and $r_{\$}(\texttt{catastrophic}) = 12$.

2.3 Standard Words

Given an infinite sequence of integers $(d_0, d_1, d_2, \ldots)$, with $d_0 \geq 0$, and $d_i > 0$ for all $i > 0$, called a *directive sequence*, define a sequence of words s_i with $i \geq 0$ of increasing length as follows: $s_0 = \texttt{b}$, $s_1 = \texttt{a}$, $s_{i+1} = s_i^{d_{i-1}} s_{i-1}$, for $i \geq 1$. Each word s_i, obtained by using the first $i-1$ integers of the directive sequence, is called *standard word* of *order* i. Without loss of generality, here we can assume that $d_0 > 0$ (otherwise, the role of b and a is exchanged.). Every standard word s_i, with $i \geq 2$, can be written as $s_i = x_i\texttt{ab}$ if i is even, $s_i = x_i\texttt{ba}$ if i is odd, where the factor x_i is a palindrome [13].

Standard words are a well studied family of binary words with a lot of interesting combinatorial properties and characterizations and appear as extreme cases in many contexts [10, 11, 22, 34, 37]. In particular, in [34], it was shown that the BWT of a binary word has exactly two runs if and only if it is a conjugate of a standard word or a conjugate of a power of a standard word.

Fibonacci words are a particular case of standard words, with directive sequence consisting of only ones. More precisely, Fibonacci words can be defined as follows: $s_0 = \texttt{b}$, $s_1 = \texttt{a}$, $s_{i+1} = s_i s_{i-1}$, for $i \geq 1$. The next few Fibonacci words are then $s_2 = \texttt{ab}$, $s_3 = \texttt{aba}$, $s_4 = \texttt{abaab}$, $s_5 = \texttt{abaababa}$, $s_6 = \texttt{abaababaabaab}$, $s_7 = \texttt{abaababaabaababaababa}$. It is easy to see that for all $i \geq 0$, $|s_i| = F_i$, where $(F_i)_{i\geq 0}$ is the *Fibonacci sequence* 1, 1, 2, 3, 5, 8, 13, 21, ..., defined by the recurrence $F_0 = F_1 = 1$, and $F_{i+1} = F_i + F_{i-1}$ for $i \geq 1$. Using the well-known fact that the Fibonacci sequence grows exponentially in i, we have that $i = \Theta(\log |s_i|)$. Moreover, for all $k \geq 1$, $s_{2k} = x_{2k}\texttt{ab}$ and $s_{2k+1} = x_{2k+1}\texttt{ba}$, where x_{2k} and x_{2k+1} are palindromes (in particular, $x_2 = \epsilon$). These words x_h, for $h \geq 2$, are also referred to as *central words*. The recursive structure of the words x_{2k} and x_{2k+1} is also known [14]: $x_{2k} = x_{2k-1}\texttt{ba}x_{2k-2} = x_{2k-2}\texttt{ab}x_{2k-1}$ and $x_{2k+1} = x_{2k}\texttt{ab}x_{2k-1} = x_{2k-1}\texttt{ba}x_{2k}$.

3 Increasing r by a $\Theta(\log n)$-Factor

In this section, we give an infinite family of words, namely the Fibonacci words of even order, on which a single edit operation, such as insertion, deletion or substitution of a character, can cause an increase of r from constant to $\Theta(\log n)$, where n is the length of the word.

To show this result, it would be sufficient to prove that $\frac{r(s')}{r(s)} = \Omega(\log n)$, where s' denotes a word obtained after a single edit operation on a word s of length n from this family. This is because the upper bound $\mathcal{O}(\log r \log n)$ [1] on the multiplicative sensitivity for this family of words is in fact $\mathcal{O}(\log n)$, since $r = 2$ for Fibonacci words. One way of proving the above lower bound is by identifying a group of rotations of s' that increase the value of r from 2 to $\Omega(\log n)$. In this section, we prove a stronger

result. In particular, we compute the explicit form of the BWT of s', where s' is obtained after an insertion, a deletion, or a substitution of a single character on a Fibonacci word of even order. These results may be of independent interest.

The impact of the three edit operations on the BWT of a Fibonacci word of even order is shown in Fig. 2.

3.1 Inserting a Character

First we recall a result from [19], namely that appending a character to the reverse of a Fibonacci word can increase the number of runs by a logarithmic factor [19]. This result was shown using the following proposition, which we report using our present terminology:

Proposition 1 ([19], Prop. 3) *Let v be the reverse of a Fibonacci word s.*

	CA	rotations of abaababaabaab	
0	7	aabaababaabab	b
1	2	aababaabaabab	b
2	10	aababaababaab	b
3	5	abaabaababaab	b
4	0	abaababaabaab	b
5	8	abaababaababa	a
6	3	ababaabaababa	a
7	11	ababaababaaba	a
8	6	baabaababaaba	a
9	1	baababaabaaba	a
10	9	baababaababaa	a
11	4	babaabaababaa	a
12	12	babaababaabaa	a

(a) Fibonacci word of order 6

	CA	rotations of abaababbaabaab	
0	8	aabaababaababb	b
1	11	aababaababbaab	b
2	2	aababbaabaabab	b
3	9	abaababaababba	a
4	0	abaababbaabaab	b
5	12	ababaababbaaba	a
6	3	ababbaabaababa	a
7	5	abbaabaababaab	b
8	7	baabaababaabab	b
9	10	baababaababbaa	a
10	1	baababbaabaaba	a
11	13	babaababbaabaa	a
12	4	babbaabaababaa	a
13	6	bbaabaababaaba	a

(b) Insertion

	CA	rotations of abaababaabaa	
0	10	aaabaababaab	b
1	7	aabaaabaabab	b
2	11	aabaababaaba	a
3	2	aababaabaaab	b
4	8	abaaabaababa	a
5	5	abaabaaabaab	b
6	0	abaababaabaa	a
7	3	ababaabaaaba	a
8	9	baaabaababaa	a
9	6	baabaaabaaba	a
10	1	baababaabaaa	a
11	4	babaabaaabaa	a

(c) Deletion

	CA	rotations of abaababaabaaa	
0	10	aaaabaababaab	b
1	11	aaabaababaaba	a
2	7	aabaaaabaabab	b
3	12	aabaababaabaa	a
4	2	aababaabaaaab	b
5	8	abaaaabaababa	a
6	5	abaabaaaabaab	b
7	0	abaababaabaaa	a
8	3	ababaabaaaaba	a
9	9	baaaabaababaa	a
10	6	baabaaaabaaba	a
11	1	baababaabaaaa	a
12	4	babaabaaaabaa	a

(d) Substitution

Fig. 2 The BWT matrix of the Fibonacci word of order 6 (a), and that of the result for 3 bit-catastrophes: (b) inserting a character in position $6 = F_{6-1} - 2$, (c) deleting the last character, (d) substituting the last character

1. *If s is of even order $2k$, for some $k \geq 1$, then* $\mathrm{BWT}(\mathtt{b}v) = \mathtt{b}^{F_{2k-2}-k+1}\mathtt{a}^{F_0}\mathtt{b}\mathtt{a}^{F_2}\mathtt{b}\mathtt{a}^{F_4}\mathtt{b}\cdots\mathtt{a}^{F_{2k-4}}\mathtt{b}\mathtt{b}\mathtt{a}^{F_{2k-2}}$. *Therefore,* $r(v\mathtt{b}) = 2k$.
2. *If s is of odd order $2k+1$, for some $k \geq 1$, then* $\mathrm{BWT}(\mathtt{a}v) = \mathtt{b}^{F_{2k-2}}\mathtt{a}\mathtt{a}\mathtt{b}^{F_{2k-4}}\mathtt{a}\mathtt{b}^{F_{2k-6}}\mathtt{a}\cdots\mathtt{b}^{F_2}\mathtt{a}\mathtt{b}^{F_0}\mathtt{a}^{F_{2k}-k+1}$. *Therefore,* $r(\mathtt{a}v) = 2k$.

This proposition implies that appending a character to the reverse of a Fibonacci word can result in a logarithmic increase in the number of runs of the BWT. To see this, recall that a well-known property of every standard word is that its reverse is one of its rotations [13]. Since s is a Fibonacci word, and thus a standard word, its reverse v has the same BWT as s. Since s is a standard word, $r(s) = 2$, and therefore, also $r(v) = 2$. Using the fact that the length of the ith Fibonacci word is F_i, and that the Fibonacci sequence $(F_i)_{i\geq 0}$ grows exponentially in i, it follows that by appending a final $\mathtt{b}$, the number of runs of the BWT is increased by a logarithmic factor in $n = |v|$, namely from $2 = \Theta(1)$ to either i (if $i = 2k$, for some k) or $i-1$ (if $i = 2k+1$, for some k), which are both $\Theta(\log n)$. Finally, since cv and vc are rotations and thus have the same BWT, appending the new character at the beginning or at the end of the word has the same effect.

Similarly to Proposition 1, we will prove that adding a character c greater than $\mathtt{b}$ and not present in the word has the same effect as adding a further $\mathtt{b}$ at the end of the reverse of a Fibonacci word of even order. Intuitively, this is because in both cases a new factor is introduced in the word, namely $\mathtt{bb}$ respectively c. Both of these factors are greater than all the other factors of the word, and they are the only changes in the word. Adding a further $\mathtt{a}$ to the reverse of a Fibonacci word of odd order, or a character smaller than $\mathtt{a}$, has a similar effect. We formalize this in the following proposition:

Proposition 2 *Let v be the reverse of the Fibonacci word s, with $|v| \geq 2$, and let c be a character different from $\mathtt{a}$ and $\mathtt{b}$.*

1. *If s is of even order $2k$ and $c > \mathtt{b}$, then $r(vc) = 2k+1$.*
2. *If s is of odd order $2k+1$ and $c < \mathtt{a}$, then $2k+2 \leq r(vc) \leq 2k+3$.*

Proof Recall that $\mathrm{BWT}(vc) = \mathrm{BWT}(cv)$. For the proof, we consider the conjugate cv.

Let us consider firstly the case in which the Fibonacci word s is of even order $2k$. Let us consider the word v' obtained by prepending the character $\mathtt{b}$ to v, i.e. $v' = \mathtt{b}v$. If we denote by $n = |v'|$, then $v = v'[1..n-1]$, and cv and v' differs only for the characters in position 0. Moreover, $cv[0] = c\mathtt{b}$ is the lexicographically greatest 2-length substring of v, and $v'[0..1] = \mathtt{bb}$ is the lexicographically greatest 2-length substring of v'. In particular, the relative lexicographic order of the conjugates $\mathrm{conj}_h(cv)$ and $\mathrm{conj}_h(v')$ with $0 \leq h \leq n-1$ is the same. Therefore the BWT of cv and v' differs only by the character preceding the conjugates starting in position 1. Note that $\mathrm{conj}_1(v') = \mathtt{b}\mathtt{a}x_{2k}\mathtt{b}$. As shown in [19, Figure 2], this is the first rotation in the bottom part of the BWT matrix of v'. By [19, Proposition 4], the character preceding $\mathrm{conj}_1(v')$ is the last character of a run of $\mathtt{b}$'s, therefore the character c which precedes $\mathrm{conj}_1(cv)$ adds only one run, namely the 1-length run of c. Since the $r(v') = 2k$, then $\mathrm{BWT}(cv)$ has the same $2k$ runs, plus the further run consisting of the single c. Therefore $r(vc) = r(cv) = 2k+1$.

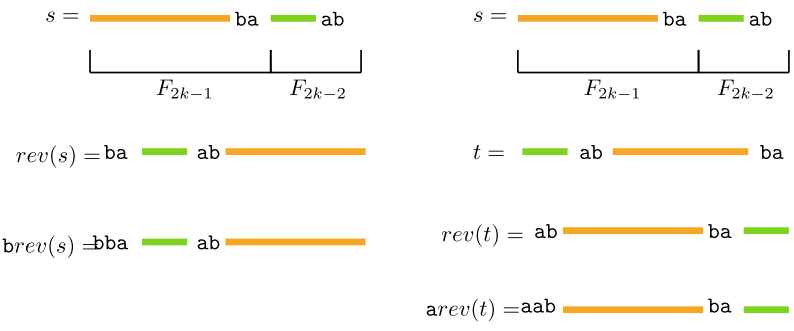
Let us consider now the case in which s is of odd order $2k+1$. Let us consider the word $v' = \mathtt{a}v$ of length n. If $k = 1$, it is easy to verify that $r(\mathtt{abaa}) = 2$ and $r(\mathtt{abac}) = 4$, so $r(vc) = 2k+2$. Let us suppose $k > 1$. From [19, Proposition 3], we know that $\mathrm{BWT}(\mathtt{a}v) = \mathtt{b}^{F_{2k-2}}\mathtt{aab}^{F_{2k-4}}\cdots\mathtt{b}^{F_2}\mathtt{ab}^{F_0}\mathtt{a}^{F_{2k}-k+1}$. Let us consider also the conjugate cv having the same BWT as vc. Since $v = v'[1..n-1]$, then cv and v' differs only for the characters in position 0. Note that $c\mathtt{a}$ and $\mathtt{aaa}$ are the lexicographically smallest substrings of cv and v' having length 2 and 3, respectively. Moreover, each of the two words occurs only once in cv and v', respectively. It follows that the relative lexicographic order of the conjugates $\mathrm{conj}_h(cv)$ and $\mathrm{conj}_h(v')$ with $1 \le h \le n-1$ is the same. The additional rotation cv starts with c, and it is now the smallest among all rotations, therefore it will be at the very beginning of the matrix. Since cv ends with $\mathtt{a}$ and the lexicographically following rotations end with $\mathtt{b}$, it increments the number of runs by 1 with respect to the BWT of v'. On the other hand, the rotation $\mathrm{conj}_0(v')$ ends with $\mathtt{a}$ and follows all the rotations that start with $\mathtt{aa}$ and end with $\mathtt{b}$, and precedes the rotation $\mathrm{conj}_{n-3}(v')$ that starts with $\mathtt{abaaa}$ and ends with $\mathtt{a}$. Additionally, the rotation ending with c contributes to r with at most 2 more runs. This is because it either falls in between runs of two distinct characters, or within a run of a single character. In total, $\mathrm{BWT}(cv)$ has at most $2k+3$ runs, and $2k+2 \le r(cv) = r(vc) \le 2k+3$. □

The following proposition can be deduced from Proposition 1 and shows that there exists at least one position in a Fibonacci word of even order where adding a character causes a logarithmic increment of r. A second position is discussed in Remark 1. These two positions are shown in Fig. 3.

Proposition 3 *Let s be the Fibonacci word of even order $2k$, and $n = |s|$. Let s' be the word resulting from inserting a $\mathtt{b}$ at position $F_{2k-1} - 2$. Then*

$$\mathrm{BWT}(s') = \mathtt{b}^{F_{2k-2}-k+1}\mathtt{a}^{F_0}\mathtt{ba}^{F_2}\mathtt{ba}^{F_4}\mathtt{b}\cdots\mathtt{a}^{F_{2k-4}}\mathtt{bba}^{F_{2k-2}},$$

and it has $2k$ runs.



(a) Insert a $\mathtt{b}$ in position $F_{2k-1} - 2$. (b) Insert an $\mathtt{a}$ in position $F_{2k} - 2$

Fig. 3 Example of adding a $\mathtt{b}$ and an $\mathtt{a}$ within the Fibonacci word of even order at positions $F_{2k-1} - 2$ and $F_{2k} - 2$, respectively. It causes a logarithmic increment of the number of BWT-runs

Proof It is known that each standard word and its reverse are conjugate. Let us consider $s = x_{2k}\mathtt{ab}$, where x_{2k} is palindrome. Moreover, from known properties on central words, $x_{2k} = x_{2k-1}\mathtt{ba}x_{2k-2} = x_{2k-2}\mathtt{ab}x_{2k-1}$ with x_{2k-1} and x_{2k-2} palindrome, as well. Let $v = \mathtt{ba}x_{2k}$ be the reverse of s. From Proposition 1, we know that $r(v\mathtt{b}) = 2k$. Since x_{2k}, x_{2k-1} and x_{2k-2} are palindromes, it follows that $s = x_{2k-1}\mathtt{ba}x_{2k-2}\mathtt{ab}$ and $v = \mathtt{ba}x_{2k-2}\mathtt{ab}x_{2k-1}$ are conjugates. In particular, $v = \mathrm{conj}_{F_{2k-1}-2}(s)$. This means that appending a $\mathtt{b}$ to v is equivalent to inserting $\mathtt{b}$ at position $F_{2k-1} - 2$ of s. □

Remark 1 Let s'' be the word resulting from inserting an $\mathtt{a}$ at position $F_{2k} - 2$. Also in this case $\mathrm{BWT}(s'')$ has $\Theta(\log n)$ runs, where $n = |s|$. To see this, let us consider the word $t = x_{2k}\mathtt{ba}$, a standard word [13, Theorem 1.1] and a conjugate of both v and s. In fact, $x_{2k}\mathtt{ba} = x_{2k-1}\mathtt{ba}x_{2k-2}\mathtt{ba} = \mathrm{conj}_{F_{2k-1}}(s)$. Since s is obtained by using a directive sequence in which the first 2k-1 integers $(d_0, d_1, \ldots, d_{2k-2})$ are all equal to 1, from [13, Proposition 10] it follows that t can be constructed by using the $2k - 2$ integers $(d_0, d_1, \ldots d_{2k-4}, d_{2k-3} + 1)$. Therefore, t is a standard word of odd order $2k - 1$. It is easy to verify that $\mathrm{rev}(t) = \mathrm{conj}_{F_{2k}-2}(s)$. By [19, Proposition 8], $r(\mathrm{rev}(t)\mathtt{a}) = 2k - 2 = \Theta(\log n)$, and by similar considerations as above, appending an $\mathtt{a}$ to $\mathrm{rev}(t)$ is equivalent to inserting $\mathtt{a}$ at position $F_{2k} - 2$ in s.

An analogous result to Proposition 3 can be proved for Fibonacci words of odd order.

3.2 Deleting a Character

We next show that deleting a character can result in a logarithmic increment in r. In particular, we consider a Fibonacci word of even order and compute the form of its BWT, as shown in the following proposition.

Proposition 4 *Let s be the Fibonacci word of even order $2k > 4$ and $\widehat{s} = s[0..n-2]$, where $n = |s|$. Then $\mathrm{BWT}(\widehat{s})$ has the following form:*

$$\mathrm{BWT}(\widehat{s}) = \mathtt{b}^{k-1}\mathtt{ab}^{F_{2k-3}-k+1}\mathtt{ab}^{F_{2k-5}}\cdots\mathtt{b}^{F_5}\mathtt{ab}^{F_3}\mathtt{aba}^{F_{2k-1}-k+1}.$$

Therefore, $\mathrm{BWT}(\widehat{s})$ has $2k$ runs.

To give the proof, we divide the BWT matrix of the word $\widehat{s}$ in three parts: *top*, *middle* and *bottom part*, showing the form of each part separately:

$$\begin{aligned}
\mathrm{BWT}_{\mathrm{top}}(\widehat{s}) &= \mathtt{b}^{k-1}\mathtt{ab}^{F_{2k-3}-k+1}, \text{ consisting of 3 runs,}\\
\mathrm{BWT}_{\mathrm{mid}}(\widehat{s}) &= \mathtt{ab}^{F_{2k-5}}\mathtt{ab}^{F_{2k-7}}\cdots\mathtt{ab}, \text{ consisting of } 2(k-2) \text{ runs,}\\
\mathrm{BWT}_{\mathrm{bot}}(\widehat{s}) &= \mathtt{a}^{F_{2k-1}-k+1}, \text{ consisting of 1 run.}
\end{aligned}$$

Altogether, we then have $3 + 2(k-2) + 1 = 2k = \Theta(\log n)$ runs.

In order to describe the structure of the matrix, we start with the following lemma that provides information on the structure of s.

Lemma 5 *Let s be the Fibonacci word of even order $2k > 4$ and $n = |s|$. Then s can be factorized as follows:*

$$\begin{aligned} s &= x_{2k-1}\mathtt{ba}x_{2k-3}\mathtt{ba}\cdots x_7\mathtt{ba}x_5\mathtt{ba}s_4 \\ &= x_{2k-2}\mathtt{ab}x_{2k-3}\mathtt{ba}x_{2k-4}\mathtt{ab}x_{2k-5}\mathtt{ba}\cdots x_4\mathtt{ab}x_3\mathtt{ba}s_4, \end{aligned}$$

where x_i denotes the central word of order i, with $3 \le i \le 2k-1$ and $s_4 = x_4\mathtt{ab} = \mathtt{abaab}$ is the Fibonacci word of order 4.

Proof It follows by using the induction on k and the fact that $s = s_{2k-1}s_{2k-2}$ and $s = x_{2k}\mathtt{ab}$, where $x_{2k} = x_{2k-1}\mathtt{ba}x_{2k-2} = x_{2k-2}\mathtt{ab}x_{2k-1}$ and $x_{2k-1} = x_{2k-3}\mathtt{ba}x_{2k-2} = x_{2k-2}\mathtt{ab}x_{2k-3}$. In fact, the equality $s_6 = x_5\mathtt{ba}x_4\mathtt{ab}$ holds since $s_4 = x_4\mathtt{ab}$. On the other hand, $s_6 = x_4\mathtt{ab}x_5\mathtt{ab} = x_4\mathtt{ab}x_3\mathtt{ba}x_4\mathtt{ab}$. Since $s_{2k+2} = s_{2k+1}s_{2k}$ and $s_{2k+1} = x_{2k+1}\mathtt{ba}$, we have that $s_{2k+2} = x_{2k+1}\mathtt{ba}x_{2k-1}\mathtt{ba}x_{2k-3}\mathtt{ba}\cdots x_7\mathtt{ba}x_5\mathtt{ba}s_4b$. On the other hand, $s_{2k+2} = s_{2k+1}x_{2k-2}\mathtt{ab}x_{2k-3}\mathtt{ba}x_{2k-4}\mathtt{ab}\cdots x_4\mathtt{ab}x_3\mathtt{ba}s_4$. The thesis follows from the fact that $s_{2k+1} = x_{2k+1}\mathtt{ba} = x_{2k}\mathtt{ab}x_{2k-1}\mathtt{ba}$. □

We identify the following 3 conjugates of the word $\widehat{s} = s[0..n-2]$ of length $n-1$ that delimit the 3 parts of the BWT matrix of the word:

$$\begin{aligned} \mathrm{conj}_{n-3}(\widehat{s}) &= \mathtt{aa}x_{2k-1}\mathtt{ba}x_{2k-3}\mathtt{ba}\cdots x_5\mathtt{baab}, \\ \mathrm{conj}_{n-5}(\widehat{s}) &= x_4\mathtt{a}x_{2k-1}\cdots x_5\mathtt{ba} \\ \mathrm{conj}_0(\widehat{s}) &= x_{2k}\mathtt{a} \end{aligned}$$

The structure of these 3 conjugates follows from Lemma 5. It is easy to see that $\mathrm{conj}_{n-3}(\widehat{s}) < \mathrm{conj}_{n-5}(\widehat{s}) < \mathrm{conj}_0(\widehat{s})$. The rotation $\mathrm{conj}_{n-3}(\widehat{s})$, starting with $\mathtt{aa}x_{2k-1}$ is the smallest rotation in the matrix due to the unique $\mathtt{aaa}$ prefix. The rotation $\mathrm{conj}_{n-5}(\widehat{s})$, starting with $x_4 = \mathtt{aba}$ indicates the beginning of the middle part, and it is the smallest rotation starting with $\mathtt{ab}$. Finally, the word itself $\widehat{s} = x_{2k}\mathtt{a}$ determines the beginning of the bottom part, namely the last long run of $\mathtt{a}$'s in the BWT.

The top part of the matrix consists of all rotations of the word starting with $\mathtt{aa}$. We give first the following lemma characterizing all occurrences of the factor $\mathtt{aa}$ in $\widehat{s}$.

Lemma 6 *Let $\widehat{s} = s[0..n-2]$, where s is the Fibonacci word of order $2k > 4$ and $n = |s|$. For every $2 \le h \le k-1$ there is exactly one occurrence of $x_{2h}\mathtt{aab}$ in $\widehat{s}$, and it is preceded by* $\mathtt{a}$.

Proof By using Lemma 5, we have that $\widehat{s} = x_{2k-1}\mathtt{ba}x_{2k-3}\mathtt{ba}\cdots x_7\mathtt{ba}x_5\mathtt{ba}x_4\mathtt{a}$. From the structural properties of the central words, i.e. $x_{2h} = x_{2h-1}\mathtt{ba}x_{2h-2}$, it follows that the word $x_{2h}\mathtt{a}$ appears as a suffix of $\widehat{s}$, for every $h = 2, \ldots, k-1$. Since $\widehat{s}$ starts with $\mathtt{ab}$, each x_{2h} ends with $\mathtt{a}$ and the factor $\mathtt{aaa}$ occurs only at position $n-3$ in $\widehat{s}$, we can conclude that the word $x_{2h}\mathtt{aab}$ occurs only once as a circular factor of $\widehat{s}$ and it is preceded by $\mathtt{a}$. □

We are now going to show that only one of the rotations of $\widehat{s}$ starting with $\mathtt{aa}$ ends with an $\mathtt{a}$, and we show where the $\mathtt{a}$ in the BWT of the top part breaks the run of $\mathtt{b}$'s.

Lemma 7 (Top part) *Given $\widehat{s} = s[0..n-2]$, where s is the Fibonacci word of order $2k > 4$ and $n = |s|$, then the first k rotations in the BWT matrix are $\mathtt{aaa}\cdots\mathtt{b} < \mathtt{a}x_4\mathtt{aab}\cdots\mathtt{b} < \mathtt{a}x_6\mathtt{aab}\cdots\mathtt{b} < \ldots < \mathtt{a}x_{2k-2}\mathtt{aab}\cdots\mathtt{b} < \mathtt{a}x_{2k}$. All other $F_{2k-3} - k + 1$ rotations starting with $\mathtt{aa}$ end with a $\mathtt{b}$.*

Proof There are $F_{2k-3} + 1$ occurrences of $\mathtt{aa}$. In fact, $\widehat{s}$ has F_{2k-1} occurrences of $\mathtt{a}$'s and $F_{2k-2} - 1$ occurrences of $\mathtt{b}$'s. Since $\mathtt{bb}$ does not occur in $\widehat{s}$, it follows that $F_{2k-2} - 1$ $\mathtt{a}$'s are followed by a $\mathtt{b}$. Therefore there are $F_{2k-1} - F_{2k-2} + 1 = F_{2k-3} + 1$ occurrences of $\mathtt{a}$ followed by an $\mathtt{a}$.

Among all the rotations that begin with $\mathtt{aa}$, the smallest one is the one starting with $\mathtt{aaa}$, which is $conj_{n-3}(\widehat{s})$. All the subsequent F_{2k-3} rotations begin with $\mathtt{aab}$. The next $k-2$ smallest rotations in this group start with $\mathtt{a}x_{2h}\mathtt{aab}$, with h ranging from 2 to $k-1$ in increasing order. In fact, from the recursive construction of s, it follows that all the occurrences of $\mathtt{aab}$ are occurrences of $\mathtt{a}x_{2h}$ for some $2 \le h \le k$. This is because each circular occurrence of $\mathtt{aab}$ is generated whenever we create $s_{2h} = s_{2h-1}s_{2h-2} = x_{2h-1}\mathtt{ba}x_{2h-2}\mathtt{ab} = x_{2h-2}\mathtt{ab}x_{2h-1}\mathtt{ab}$. Moreover, each central word x_i starts with $\mathtt{ab}$ and ends with $\mathtt{ba}$. The occurrences of $\mathtt{a}x_{2h}$ in $\widehat{s}$, with $2 \le h \le k-1$ can be followed by $\mathtt{ab}$, $\mathtt{ba}$ or $\mathtt{aab}$. By using Lemma 6, the occurrences of $\mathtt{a}x_{2h}\mathtt{aab}$ are all distinct and each of these factors appears only once. This means that the next $k-2$ smallest rotations start with $\mathtt{a}x_{2h}\mathtt{aab}$, for h from 2 to $k-1$ in increasing order. Finally, the next smallest rotation is $\mathtt{a}x_{2k}$. Note that only $\mathtt{a}x_{2k}$ is preceded by an $\mathtt{a}$, therefore the k smallest rotations of $\widehat{s}$ are all preceded by a $\mathtt{b}$ except for the largest of them which is preceded by an $\mathtt{a}$. This shows that the k smallest rotations in the BWT matrix form two runs: $\mathtt{b}^{k-1}\mathtt{a}$.

All the remaining $F_{2k-3} - k + 1$ rotations starting with $\mathtt{aa}$ correspond to some occurrence of $\mathtt{a}x_{2h}\mathtt{ba}$ for some $2 \le h < k$. In fact, all the occurrences of $\mathtt{a}x_{2h}\mathtt{ab}$ are also occurrences of either $\mathtt{a}x_{2k}$ or $\mathtt{a}x_{2t}\mathtt{aab}$ for some $t > h$. Then the correspondent rotations are lexicographically greater than $\mathtt{a}x_{2k}$, which is prefixed by $\mathtt{a}x_{2h}\mathtt{ab}$. Since there is a unique occurrence of $\mathtt{aaa}$, all these rotations starting with $\mathtt{aa}$ are preceded by $\mathtt{b}$ by construction. Therefore, we have $\mathtt{b}^{k-1}\mathtt{ab}^{F_{2k-3}-k+1}$ in the top part of the BWT matrix of $\widehat{s}$. □

Figure 4 displays the structure of the middle part of the BWT of $\widehat{s}$. To determine the number of runs in this middle part of the matrix, it is crucial first to consider that all the rotations starting with x_h, with $h = 4, \ldots, 2k-1$, are grouped in the BWT matrix. Specifically, these occur in blocks where rotations starting with $x_h\mathtt{a}$, h odd, are immediately preceded by the unique rotation starting with $x_{h-1}\mathtt{aa}$, and immediately followed by the rotation starting with $x_{h+1}\mathtt{aa}$, as illustrated in Fig. 4. This is proved in the following lemma.

Lemma 8 *For $4 \le h \le 2k-1$, rotations starting with some $x_h\mathtt{aa}$ are smaller then rotations starting with $x_{h+1}\mathtt{a}$. In particular, the word $\widehat{s} = x_{2k}\mathtt{a}$ is greater than any of the rotations above. Moreover, if h is odd, rotations starting with some $x_h\mathtt{a}$ are smaller then rotations starting with x_{h+1}.*

Proof Every x_h is a prefix of x_{h+1}. Since there is exactly one circular occurrence of $\mathtt{aaa}$ in $\widehat{s}$, then $x_{h+1}\mathtt{a}$ is either prefixed by $x_h\mathtt{ab}$ or by $x_h\mathtt{ba}$, i.e. the $\mathtt{aaa}$ factor

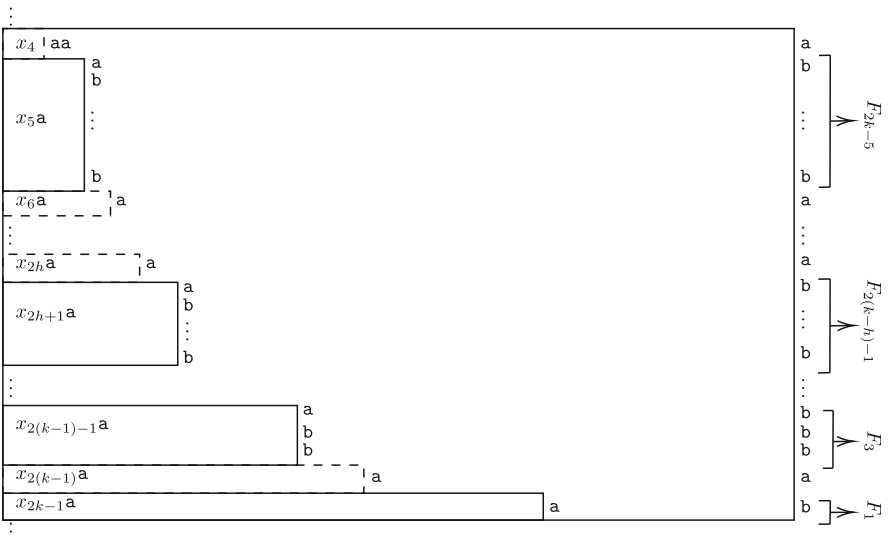


Fig. 4 The middle part $BWT_{mid}(\widehat{s})$ of the BWT matrix for the deletion of the last character of a Fibonacci word of even order $2k$ is shown

occurs earlier in $x_h\mathtt{aa}$. In both cases, the first claim holds. Finally, the thesis follows by observing that if h is odd, $x_h\mathtt{ba}$ is a prefix of x_{h+1}. □

Lemma 9 *Let x_h be a central word in a Fibonacci word. Then x_h appears only twice in x_{h+1}, once as a prefix and once as a suffix.*

Proof Let us suppose, by contradiction, that there exists an h such that x_h has another occurrence in x_{h+1} that is neither a prefix nor a suffix. It is easy to verify that for x_3, x_4, and x_5, the claim holds. In fact, $x_3 = \mathtt{a}$, $x_4 = \mathtt{aba}$, $x_5 = \mathtt{abaaba}$ and $x_6 = \mathtt{abaababaaba}$. Then, we can assume $h > 5$ and h is the smallest value for which the claim is not true. By using the recursive structure of x_{h+1}, it follows that $|x_{h+1}| = |x_h| + |x_{h-1}| + 2 = 2|x_{h-1}| + |x_{h-2}| + 4$. Let us denote by u and v the prefix and the suffix of x_{h+1} of length $|x_h|$, respectively. It is easy to verify that u and v must have an overlap of length $|x_{h-2}|$. Let w be the internal factor in x_{h+1} that is neither its suffix nor its prefix, but is equal to x_h. The overlap of w with both u and v has a length greater than $|x_{h-2}|$. Since $w = x_h$, it can be decomposed as $x_{h-1} \cdot cd \cdot x_{h-2} = x_{h-2} \cdot dc \cdot x_{h-1}$, with $c, d \in \{\mathtt{a}, \mathtt{b}\}$ and $c \neq d$. If the overlap with u has a length greater than $|x_{h-1}|$, then there would exist an occurrence of x_{h-1} in $u = x_h$ that is neither a prefix nor a suffix, which would contradict the minimality of h. If the overlap with u has a length smaller than or equal to $|x_{h-1}|$, then w would have an overlap with v of length greater than $2|x_{h-2}| + 2 > |x_{h+1}|$. This would imply an occurrence of x_{h-1} in $v = x_h$ that is neither a prefix nor a suffix. In this case as well, the minimality of h would be contradicted. □

Lemma 10 *For every $t \geq 4$ and $h \leq t - 1$, a central word x_h appears in x_t only as a prefix or a suffix of some occurrence of x_{h+1} in x_t.*

Proof We provide a proof by induction. For the base case, observe that the claim holds for $x_4 = \mathtt{aba}$, so for the inductive step let us assume that, for some $t \geq 4$, every occurrence of x_h within x_t occurs as prefix or suffix of a factor x_{h+1} of x_t, and let us prove the thesis holds for $t+1$. By definition, $x_{t+1} = x_t \cdot cd \cdot x_{t-1} = x_{t-1} \cdot dc \cdot x_t = x_{t-1} \cdot dc \cdot x_{t-2} \cdot cd \cdot x_{t-1}$, for some $c, d \in \{\mathtt{a}, \mathtt{b}\}$ such that $c \neq d$. By the inductive hypothesis, the thesis holds within the prefix and the suffix x_t, and by Lemma 9 these are the only two occurrences in x_{t+1}. Moreover, the inductive hypothesis also holds in x_{t-2}, which is the overlap of the two occurrences of x_t. In other words, every occurrence of a central word of order $3 \leq h \leq t-1$ that falls within an occurrence of x_t and every occurrence of a central word of order $3 \leq h' \leq t-3$ that falls within x_{t-2} is either a prefix or a suffix of x_{h+1} and $x_{h'+1}$ respectively. We now need to prove that each central word x_h overlapping with both the prefix and the suffix x_t of x_{t+1} and that is not fully contained neither in $x_{t+1}[0..|x_t|-1]$ nor in $x_{t+1}[n-|x_t|..n-1]$ is still a prefix of x_{h+1}. Since the overlap between the prefix and the suffix of size $|x_t|$ has length $|x_{t-2}|$, the only central word satisfying this hypothesis is x_{t-1}. If such an occurrence of x_{t-1} exists, then there must exist $u, v \neq \varepsilon$ such that $x_{t-1} = ux_{t-2}v$, where the factors $x_{t-2}v$ and ux_{t-2} are respectively prefix and suffix of x_t. However, this factorization contradicts that x_{t-2} occurs in x_{t-1} only as prefix and suffix (Lemma 9), therefore such an occurrence of x_{t-1} does not exist and the thesis follows. □

Corollary 11 *Let $k > 2$ and $3 \leq h \leq 2k-1$. Every occurrence of x_h in $\widehat{s} = x_{2k}\mathtt{a}$ is either a prefix or a suffix of some occurrence of x_{h+1} in $\widehat{s}$.*

Proof No central word of a Fibonacci word contains the substring $\mathtt{aaa}$, or ends with the substring $\mathtt{aa}$, i.e., no central word of order $t \leq 2k$ crosses the last $\mathtt{a}$ of $\widehat{s}\mathtt{a}$. Therefore, every occurrence of x_h in $\widehat{s}$ is a factor of x_{2k}, and the thesis follows from Lemma 10. □

Lemma 12 *There are $F_{2h}-1$ occurrences of $\mathtt{b}x_{2(k-h)}\mathtt{a}$ and F_{2h+1} occurrences of $\mathtt{b}x_{2(k-h)-1}\mathtt{a}$ as circular factors in $\widehat{s} = x_{2k}\mathtt{a}$, for every $0 \leq h \leq k-2$.*

Proof The claim can be proved by induction. For $h = 0$, the statement follows from the fact that there is one occurrence of $\mathtt{b}x_{2k}\mathtt{a}$ in the Fibonacci word of order $2k$. There cannot be any other occurrences because the word is primitive. Therefore there are $F_0 - 1 = 0$ occurrences in $\widehat{s}$ because of the missing $\mathtt{b}$ at the end.

By Lemma 9, x_{2k-1} has only two occurrences in x_{2k}. Since x_{2k-1} has $\mathtt{ab}$ as a suffix and does not have $\mathtt{aaa}$ as a factor, the only occurrences of x_{2k-1} in $\widehat{s} = x_{2k}\mathtt{a} = x_{2k-1}\mathtt{b}\mathtt{a}x_{2k-2}\mathtt{a} = x_{2k-2}\mathtt{ab}x_{2k-1}\mathtt{a}$ are exactly the occurrences of x_{2k-1} in x_{2k}, namely as a prefix and as a suffix. This follows from Corollary 11. Therefore, there are $F_1 = 1$ occurrences of $\mathtt{b}x_{2k-1}\mathtt{a}$.

Note that, for every $0 \leq h \leq k-2$, the occurrence of any $\mathtt{b}x_{2h}\mathtt{a}$ in position $F_{2k}-1$ of the Fibonacci word of order $2k$ is missing in $\widehat{s}$ due to the missing $\mathtt{b}$ at the end of the word.

Let us suppose the statement holds for all $i \leq h$. By Corollary 11, $x_{2(k-h)-2}$ appears in $\widehat{s}$ only as a prefix (followed by $\mathtt{a}$) and as a suffix (preceded by $\mathtt{a}$) of $x_{2(k-h)-1}$. Then, $\mathtt{b}x_{2(k-h)-2}\mathtt{a}$ is a prefix of $\mathtt{b}x_{2(k-h)-1}$. Therefore, the factor $\mathtt{b}x_{2(k-h)-2}\mathtt{a}$ appears as a

prefix of $\mathtt{b}x_{2(k-h)-1}\mathtt{a}$ and as a prefix of $\mathtt{b}x_{2(k-h)-1}\mathtt{b}$ (which is a prefix of $\mathtt{b}x_{2(k-h)}\mathtt{a}$). Moreover, the mentioned occurrences are distinct because $\mathtt{b}x_{2(k-h)-1}\mathtt{a}$ is not a prefix of $\mathtt{b}x_{2(k-h)}\mathtt{a}$.

Therefore, by induction, the number of occurrences of $\mathtt{b}x_{2(k-h)-2}\mathtt{a}$ is equal to the sum of the number of occurrences of $\mathtt{b}x_{2(k-h)}\mathtt{a}$ and those of $\mathtt{b}x_{2(k-h)-1}\mathtt{a}$: $F_{2h} - 1 + F_{2h+1} = F_{2h+2} - 1$.

Similarly, by Corollary 11, $x_{2(k-h)-3}$ appears appears in $\widehat{s}$ only as a prefix (followed by $\mathtt{b}$) and as a suffix (preceded by $\mathtt{b}$) of $x_{2(k-h)-2}$. Then, $\mathtt{b}x_{2(k-h)-3}\mathtt{a}$ is a suffix of $x_{2(k-h)-2}\mathtt{a}$. Therefore, the factor $\mathtt{b}x_{2(k-h)-3}\mathtt{a}$ appears as a suffix of $\mathtt{b}x_{2(k-h)-2}\mathtt{a}$ and as a suffix of $\mathtt{a}x_{2(k-h)-2}\mathtt{a}$ (which is a suffix of $\mathtt{b}x_{2(k-h)-1}\mathtt{a}$). Moreover, such occurrences are distinct because $\mathtt{b}x_{2(k-h)-2}\mathtt{a}$ is not a suffix of $\mathtt{b}x_{2(k-h)-1}\mathtt{a}$. Finally, $\mathtt{b}x_{2(k-h)-3}\mathtt{a}$ appears once also as suffix of $\mathtt{a}x_{2(k-h)-2}\mathtt{a}$, starting at the last position of $\widehat{s}$. Therefore, by induction, the number of occurrences of $\mathtt{b}x_{2(k-h)-3}\mathtt{a}$ is equal to the sum of the number of occurrences of $\mathtt{b}x_{2(k-h)-2}\mathtt{a}$ and those of $\mathtt{b}x_{2(k-h)-1}\mathtt{a}$ plus one: $F_{2h+2} - 1 + F_{2h+1} + 1 = F_{2h+3}$. □

Lemma 13 (Middle part) *The middle part contributes to $r(\hat{s})$ with $2(k-2)$ runs in the following form:* $\mathtt{ab}^{F_{2k-5}}\mathtt{ab}^{F_{2k-7}}\cdots\mathtt{ab}^{F_3}\mathtt{ab}$.

Proof From the recursive construction of s, it follows that all the occurrences of $\mathtt{ab}$ are generated as a prefix of some s_h, for some $2 \le h \le 2k$. This is because, for every $2 < h < 2k-2$, $s_{h+2} = s_{h+1}s_h$ with $s_2 = \mathtt{ab}$ and $s_1 = \mathtt{a}$. From the recursive structure of each central word x_h, it follows that all the rotations of $\widehat{s}$ starting with $\mathtt{ab}$ are prefixed either by $\mathtt{ab}x_h$ for some $4 \le h \le 2k-1$ or by a central word x_h, for some $4 \le h \le 2k$. Since $\mathtt{abaa}$ is a prefix of $\widehat{s}$, the rotations starting with $\mathtt{ab}x_h$ are in the bottom part. So here we consider the rotations that start with some x_h, with $h \ge 4$. Recall that $\widehat{s} = x_{2k-1}\mathtt{ba}x_{2k-2}\mathtt{a} = x_{2k-2}\mathtt{ab}x_{2k-1}\mathtt{a}$. Since each x_h appears only as a prefix and a suffix of every occurrence of x_{h+1} (Corollary 11), it follows that every occurrence of x_h in $\widehat{s}$, for $4 \le h < 2k-1$ can be followed by $\mathtt{aa}$, $\mathtt{ab}$, or $\mathtt{ba}$. Note that, by Corollary 11, x_{2k-1} occurs only twice in $\widehat{s}$, one occurrence is followed by $\mathtt{aa}$, the other one followed by $\mathtt{ba}$. Moreover, the rotation starting with $x_4\mathtt{aa}$ is the smallest in lexicographic order. For every $h \ge 4$, we can distinguish two cases:

1. h is even. By using Lemma 8, the rotation starting with $x_h\mathtt{aa}$ precede all the rotations starting with $x_{h+1}\mathtt{aa}$ or $x_{h+1}\mathtt{ab}$. Because of the recursive structure of the central words, $x_h\mathtt{ab}$ is a prefix of every x_t, with $t > h$. It follows that each rotation starting with $x_h\mathtt{ab}$ has also x_{h+1} as a prefix. Moreover, all the rotations starting with $x_h\mathtt{ba}$ are lexicographically greater than $\widehat{s} = x_{2k}$, and then they are in the bottom part of the BWT matrix.
2. h is odd. Because of the recursive structure of the central words, $x_h\mathtt{ba}$ is a prefix of every x_t, with $5 \le h \le 2k-1$. It follows that each rotation starting with $x_h\mathtt{ba}$ has also x_{h+1} as a prefix. Moreover, the rotations starting with $x_h\mathtt{ab}$ are lexicographically smaller than $\widehat{s}$.

It follows that all the rotations of $\widehat{s}$ in this middle part of the BWT matrix can be partitioned into $k-2$ blocks. In fact, for every $2 \le h \le k-1$, the first rotation is prefixed by $x_{2h}\mathtt{aa}$. By using Lemma 6, such a rotation ends with $\mathtt{a}$. The next rotation starts with $x_{2h+1}\mathtt{aa}$. In case $h \ne k-1$, the subsequent rotations in the block start with

$x_{2h+1}\mathtt{ab}$. If $h = k-1$, the last rotation of the block starts with $x_{2k-1}\mathtt{aa}$. Because of the recursive construction of the central words (i.e. $x_{2(h+1)} = x_{2h}\mathtt{ab}x_{2h+1}$), all the occurrences of x_{2h+1} followed by $\mathtt{a}$ must appear as a suffix of $x_{2(h+1)}$, then they end with $\mathtt{b}$. In fact, by Corollary 11, x_{2h+1} appears in $\widehat{s}$ only as a prefix and as a suffix in $x_{2(h+1)}$. By using Lemma 12 there are $F_{2(k-h)-1}$ of such rotations. The sequence of the block above described is illustrated in Fig. 4. □

The rotations that divide the middle part from the bottom part are the two rotations prefixed by the two occurrences of x_{2k-1}. By properties of Fibonacci words, one rotation is prefixed by $x_{2k-1}\mathtt{a}$ (end middle part) and the other by $x_{2k-1}\mathtt{b}$ (beginning bottom part). The latter follows the first in lexicographic order. Note that the rotation starting with $x_{2k-1}\mathtt{b}$ is $\widehat{s}$, namely $x_{2k-1}\mathtt{ba}x_{2k-2}\mathtt{a}$.

Lemma 14 (Bottom part) *All rotations greater than* $\widehat{s} = x_{2k-1}\mathtt{ba}x_{2k-2}\mathtt{a}$ *end with* $\mathtt{a}$.

Proof From Lemma 7 we have that $k-1+F_{2k-3}-k+1$ rotations ending with $\mathtt{b}$ have already appeared in the matrix, and from Lemma 13 $F_{2k-5}+\cdots+F_3+F_1$ rotations ending with $\mathtt{b}$ have already appeared in the matrix. Summing the number of $\mathtt{b}$'s we have $k-1+F_{2k-3}-k+1+F_{2k-5}+\cdots+F_3+F_1 = F_{2k-3}+F_{2k-5}+\cdots+F_3+F_1$. We can decompose each odd Fibonacci number F_{2x+1} in the sum $F_{2x}+F_{2x-1}$. Therefore, the previous sum becomes $F_{2k-4}+F_{2k-5}+F_{2k-6}+F_{2k-7}+\cdots+F_2+F_1+F_1$. For every Fibonacci number F_x, it holds that $F_x = F_{x-2}+F_{x-3}+F_{x-4}+\cdots+F_2+F_1+2$. Therefore, $F_{2k-4}+F_{2k-5}+F_{2k-6}+F_{2k-7}\cdots+F_2+F_1+F_1 = F_{2k-2}-1$, which is exactly the number of $\mathtt{b}$'s in $\widehat{s}$. Therefore all the remaining rotations end with $\mathtt{a}$. □

In the context of repetitiveness measures of words, a measure λ is called *monotone* if, for each word $v \in \Sigma^*$ and for each letter $c \in \Sigma$, it holds that $\lambda(v) \leq \lambda(vc)$. Since we have shown that appending or deleting a single character can substantially increase the parameter r, the following known result on the monotonicity of r can be derived:

Corollary 15 *The measure* r *is not monotone.*

3.3 Substituting a Character

In this subsection, we show how to increment r by a logarithmic factor by substituting a character. Consider a Fibonacci word s of even order in which we replace the last $\mathtt{b}$ by an $\mathtt{a}$. Denoting this word by s', we will prove that $\mathrm{BWT}(s')$ has $\Theta(\log n)$ runs, where n is the length of the word. We start with the following lemma in which we assess how the BWT changes when we append or prepend to a Lyndon word a character that is smaller than or equal to the smallest character appearing in the word itself.

Lemma 16 *Let* $v \in \Sigma^*$ *be a Lyndon word containing at least two distinct letters and let* $c \in \Sigma$ *be smaller than or equal to the smallest character occurring in* v*, and let* $n = |v|$*. Then,* $\mathrm{BWT}(cv) = \mathrm{BWT}(v)[0] \cdot c \cdot \mathrm{BWT}(v)[1..n-1]$*. Therefore,* $r(v) \leq r(cv) = r(vc) \leq r(v)+2$.

Proof We can obtain the lexicographic order of the rotations of cv, or equivalently vc, from the order of the rotations of v. To do so, we show that given two rotations

$conj_i(v) < conj_j(v)$, with $i \neq j$, if $conj_i(v) < conj_j(v)$ then $v[i..|v|-1]cv[0..i-1] < v[j..|v|-1]cv[0..j-1]$.

Note that v is the smallest rotation in its BWT matrix, and that $c \neq v[n-1]$ (otherwise $conj_{n-1}(v) < v$, a contradiction). Let us denote by $conj_h(v)$, for some h, the second rotation in the BWT matrix. Since v is primitive, there exists a unique circular factor u smaller than all the other circular factors having the same length. In fact, if $t = |\text{lcp}(v, conj_h(v))|$, then $u = v[0..t]$. Moreover, for all $\ell < |u|$, $u[0..\ell-1]$ is the smallest circular factor of length ℓ occurring in v. We can then distinguish two cases.

The first case is when $|\text{lcp}(conj_i(v), conj_j(v))| < \min\{|v|-i+1, |v|-j+1\}$. Under this condition, it follows that the insertion of the c does not affect the relative order between $v[i..|v|-1]cv[0..i-1]$ and $v[j..|v|-1]cv[0..j-1]$.

Otherwise, if $|\text{lcp}(conj_i(v), conj_j(v))| \geq \min\{|v|-i+1, |v|-j+1\}$, note that $i > j$, i.e. $|v[i..|v|-1]| < |v[j..|v|-1]|$. This follows by observing that both $v[i..|v|-1]$ and $v[j..|v|-1]$ are (circularly) followed by u that is unique, and by contradiction if $i < j$ then u would circularly occur before in $conj_j(v)$ with respect to $conj_i(v)$, which contradicts $conj_i(v) < conj_j(v)$. We can now further distinguish between two subcases: when either (i) u is a prefix of $v[0..i-1]$ or (ii) $v[0..i-1]$ is a proper prefix of u.

For the subcase (i), as $|\text{lcp}(conj_i(v), conj_j(v))| \geq |v[i..|v|-1]|$, and the factor u is a prefix of $v[0..i-1]$, the first distinct character between $conj_i(v)$ and $conj_j(v)$ lies within the unique occurrence of u in $conj_i(v)$. After the letter c is inserted, $conj_i(v)$ becomes $v[i..|v|-1]cv[0..i-1]$, yielding a factor cu occurring at position $|v|-i+1$ that is also unique and smallest among all the factors of length $|cu|$ in cv. Whatever factor appears in $v[j..|v|-1]cu$ at position $|v|-i+1$, has to be greater than cu, and the order is preserved. For the subcase (ii), recall that since $v[0..i-1]$ is a proper prefix of u, $v[0..i-1]$ is also the smallest i-length circular factor in lexicographical order occurring in v, but differently from u the circular factor $v[0..i-1]$ is repeated (otherwise $|u| \leq |v[0..i-1]|$, contradiction). By primitivity of v, the first distinct character between $conj_i(v)$ and $conj_j(v)$ lies within $v[0..i-1]$, i.e., within $conj_i(v)[|v|-i+1..|v|-1]$. After the insertion of the symbol c the analogous behavior of subcase (i) is observed.

We conclude the proof by observing that, with respect to the original BWT, we have one extra rotation, and one rotation for which the letter in the BWT has changed, which are cv and vc respectively. Observe that by construction, cv is now the smallest among all of its rotations, which ends with the last letter of v. On the other hand, vc is now the second smallest rotation and it ends with c. Hence, $\text{BWT}(cv) = \text{BWT}(v)[0] \cdot c \cdot \text{BWT}(v)[1..|v|-1]$.

Finally, from the structure of $\text{BWT}(cv)$ it follows that:

- if $c = \text{BWT}(v)[1]$, then $r(cv) = r(v)$;
- if $c \neq \text{BWT}(v)[1]$ and $\text{BWT}(v)[0] \neq \text{BWT}(v)[1]$, then $r(cv) = r(v)+1$;
- if $c \neq \text{BWT}(v)[1]$ and $\text{BWT}(v)[0] = \text{BWT}(v)[1]$, then $r(cv) = r(v)+2$.

□

Proposition 17 *Let s be the Fibonacci word of even order $2k > 4$. Let s' be the word resulting from substituting in s a* b *with an* a *at position $F_{2k}-1$. Then* $\text{BWT}(s')$ *has*

the following form:

$$\mathrm{BWT}(s') = \mathtt{bab}^{k-2}\mathtt{ab}^{F_{2k-3}-k+1}\mathtt{ab}^{F_{2k-5}}\cdots\mathtt{b}^{F_5}\mathtt{ab}^{F_3}\mathtt{aba}^{F_{2k-1}-k+1}.$$

Therefore, $\mathrm{BWT}(s')$ *has* $2k+2$ *runs.*

Proof Let $n = |s'|$. Observe that $s' = \widehat{s}\mathtt{a} = \widehat{s}[0..n-4]\mathtt{a}\widehat{s}[n-3..n-2]$. From the proof of Lemma 7, we know that $conj_{n-3}(\widehat{s})$ is the Lyndon rotation of $\widehat{s}$. Since $\mathrm{BWT}(s') = \mathrm{BWT}(\widehat{s}[0..n-4]\mathtt{a}\widehat{s}[n-3..n-2]) = \mathrm{BWT}(\mathtt{a}conj_{n-3}(\widehat{s}))$, the thesis follows by Proposition 4 and Lemma 16. □

4 Additive $\Theta(\sqrt{n})$ Factor

In the previous section, we proved that a single edit operation can cause a multiplicative increase by a logarithmic factor in the number of runs. In this section, we will exhibit an infinite family of words on which a single edit operation can cause an additive increment of r by $\Theta(\sqrt{n})$ (see Def. 19 below). More precisely, for this family of words, we explicitly compute the BWT after applying an insertion, deletion, or substitution of a single character. These results may be of independent interest.

As we saw in the previous section, there exist infinite families of words such that $r = \Theta(\log n)$, where n is the length of the word. Other families with a logarithmic number of runs of the BWT are also known from the literature, e.g. the Thue-Morse words [8, 16]. Moreover, there exist words such that r is maximal, i.e., $r(w) = |w|$. For instance, if $w = \mathtt{aaaabbababbbbaabab}$, then $r(w) = 18 = |w|$ [33]. Next, we show that there is no gap between these two scenarios, i.e., it is possible to construct infinite families of words w such that $r(w) = \Theta(n^{1/k})$, for any $k > 1$.

Proposition 18 *Let* k *be a positive integer. There exists an infinite family* T_k *of binary words such that* $r(w) = \Theta(n^{1/k})$*, for any* $w \in T_k$*.*

Proof We can define the set $T_k = \{w_{i,k} = \prod_{j=1}^{i} \mathtt{ab}^{j^k} \mid i \geq 1\}$. We can state that $|w_{i,k}| = \Theta(i^{k+1})$. Moreover, $r(w_{i,k}) = \Theta(i)$. In fact, each $\mathtt{a}$ in $\mathrm{BWT}(w_{i,k})$ corresponds to the last letter of one of the rotations having prefix $\mathtt{b}^{j^k}\mathtt{a}$, for some $1 \leq j \leq i$. On the other hand, all the rotations with prefix $\mathtt{ab}$, as well as the remaining rotations with prefix $\mathtt{b}^{\ell}\mathtt{a}$ for all $1 \leq \ell \leq i^k$, end with $\mathtt{b}$. It follows then that whenever $k \geq 2$, all the $\mathtt{a}$'s in $\mathrm{BWT}(w_{i,k})$ are separated by an equal-letter run of $\mathtt{b}'s$, leading to $r(w_{i,k}) = 2i = \Theta(i)$. However, note that for a fixed ℓ, the rotations starting with $b^{\ell}\mathtt{a}$ are sorted according to the length of the maximal run of $\mathtt{b}$'s following the common prefix. Thus, even for $k = 1$, there is only one run of consecutive $\mathtt{a}$'s in $\mathrm{BWT}(w_{i,k})$, while the remaining are separated. More in detail, $\mathrm{BWT}(w_{i,1}) = \mathtt{bb}^{i}\mathtt{ab}^{i-1}\mathtt{a}\cdots\mathtt{ab}^{3}\mathtt{ab}^{2}\mathtt{aa}$. Hence, $r(w_{i,1}) = 2i - 2 = \Theta(i)$. The claim follows by observing that for the family of words $w_{i,k}$ it holds that $r(w_{i,k}) = \Theta(i) = \Theta(n^{\frac{1}{k+1}})$, where $n = |w_{i,k}|$. □

We will show that the following family of words satisfies that a single edit operation can cause an additive increment of r by $\Theta(\sqrt{n})$.

Definition 19 For any $k > 5$, let $s_i = \mathtt{ab}^i\mathtt{aa}$ and $e_i = \mathtt{ab}^i\mathtt{aba}^{i-2}$ for all $2 \leq i \leq k-1$, and $q_k = \mathtt{ab}^k\mathtt{a}$. Then,

$$w_k = (\prod_{i=2}^{k-1} s_i e_i) q_k = (\prod_{i=2}^{k-1} \mathtt{ab}^i\mathtt{aaab}^i\mathtt{aba}^{i-2})\mathtt{ab}^k\mathtt{a}.$$

The length of these words can be easily deduced from their definition.

Remark 2 Let $n = |w_k|$ for some $k > 5$. It holds that $n = \sum_{i=2}^{k-1}(3i+4) + (k+2) = (3k^2 + 7k - 18)/2$. Moreover, it holds that $k = \Theta(\sqrt{n})$.

The following lemma will be used to show how the rotations of w_k can be sorted according to the factorization $s_2 e_2 \cdots s_{k-1} e_{k-1} q_k$.

Lemma 20 *Let $k > 5$ be an integer. Then, $s_2 < e_2 < s_3 < e_3 < \ldots < s_{k-1} < e_{k-1} < q_k$. Moreover the set $\mathcal{U} = \bigcup_{i=2}^{k-1}\{s_i, e_i\} \cup \{q_k\}$ is prefix-free.*

Proof For the first claim, note from the definition of the words e_i, s_i and q_k that for $i \in [2, k-1]$ it holds $s_i < e_i$, for $i \in [2, k-2]$ it holds $e_i < s_{i+1}$, and it holds $e_{k-1} < q_k$. For the second claim, observe that for any two distinct strings x and y in the set $\mathcal{U}$ starting with $\mathtt{ab}^j\mathtt{a}$ and $\mathtt{ab}^{j'}\mathtt{a}$ respectively, there are two possible cases. If $j = j'$ then x and y are s_i and e_i respectively, and none of them is a prefix of the other. Otherwise, w.l.o.g. $j < j'$, so $x = \mathtt{ab}^j\mathtt{a}x'$ and $y = \mathtt{ab}^j\mathtt{b}y'$ for some x' and y'. Hence $x[j+2] \neq y[j+2]$ and none of them is a prefix of the other. Thus, the set $\mathcal{U}$ is prefix-free. □

4.1 Characterizing the BWT of w_k

In order to characterize the BWT of the word

$$w_k = (\prod_{i=2}^{k-1} s_i e_i) q_k = (\prod_{i=2}^{k-1} \mathtt{ab}^i\mathtt{aa} \cdot \mathtt{ab}^i\mathtt{aba}^{i-2}) \cdot \mathtt{ab}^k\mathtt{a},$$

we divide its BWT matrix into disjoint ranges of consecutive rotations sharing the same (specific) prefixes, and characterize the substring of BWT(w_k) corresponding to each one of these prefixes.

Definition 21 Given $x, w \in \Sigma^*$, we denote by $\beta(x, w)$ the substring of BWT(w) corresponding to the range of contiguous rotations prefixed by x. We omit the second parameter of $\beta(x, w)$ when it is clear from the context.

The structure of the whole BWT matrix of w_k is summarized in Table 1. The following series of lemmas characterize the substring of BWT(w_k) corresponding to each range to be considered.

Table 1 Scheme of the BWT matrix of a word w_k with $k > 5$

Block prefix	Ordering factor	BWT
$\mathtt{a}^{k-2}\mathtt{b}$	$\mathtt{b}^{k-1}\mathtt{a}$	b
$\mathtt{a}^{k-3}\mathtt{b}$	$\mathtt{b}^{k-2}\mathtt{aa}$	b
	$\mathtt{b}^{k-1}\mathtt{a}$	a
⋮	⋮	⋮
$\mathtt{a}^{4}\mathtt{b}$	$\mathtt{b}^{5}\mathtt{aa}$	b
	$\mathtt{b}^{6}\mathtt{aa}$	a
	⋮	⋮
	$\mathtt{b}^{k-1}\mathtt{a}$	a
aaab	bab	b
	bbaba	b
	bbbabaa	b
	bbbbaa	b
	bbbbabaaa	b
	bbbbbaa	a
	bbbbbabaaaa	b
	⋮	⋮
	$\mathtt{b}^{k-2}\mathtt{aa}$	a
	$\mathtt{b}^{k-2}\mathtt{aba}^{k-3}$	b
	$\mathtt{b}^{k-1}\mathtt{a}$	a

Block prefix	Ordering factor	BWT
	baa	b
	bab	a
	bbaba	a
	bbbaa	b
	bbbabaa	b
aab	bbbbaa	a
	bbbbabaaa	a
	⋮	⋮
	$\mathtt{b}^{k-2}\mathtt{aa}$	a
	$\mathtt{b}^{k-2}\mathtt{aba}^{k-3}$	a
	$\mathtt{b}^{k-1}\mathtt{a}$	a
	$\mathtt{a}^{k-3}q_k$	b
	$\mathtt{a}^{k-4}s_{k-1}$	b
	⋮	⋮
	s_3	b
ab	baa	a
	bab	b
	bbaa	b
	bbaba	a
	bbbaa	a
	bbbabaa	a
	⋮	⋮
	$\mathtt{b}^{k-1}\mathtt{a}$	a

Block prefix	Ordering factor	BWT
	$\mathtt{a}^{k-4}q_k$	a
	$\mathtt{a}^{k-5}s_{k-1}$	a
	⋮	⋮
	$\mathtt{a}^{2}s_6$	a
	$\mathtt{a}e_2$	b
	$\mathtt{a}e_3$	b
	$\mathtt{a}e_4$	a
	$\mathtt{a}s_5$	a
ba	$\mathtt{a}e_5$	b
	$\mathtt{a}e_6$	b
	⋮	⋮
	$\mathtt{a}e_{k-1}$	b
	s_2	b
	s_4	a
	$\mathtt{ba}^{k-3}q_k$	b
	$\mathtt{ba}^{k-4}s_{k-1}$	b
	⋮	⋮
	$\mathtt{b}s_3$	b
	bbbaa	a

Block prefix	Ordering factor	BWT
	$\mathtt{a}e_2$	a
	$\mathtt{a}e_3$	b
	⋮	⋮
	$\mathtt{a}e_{k-1}$	b
bba	s_2	b
	$\mathtt{ba}^{k-3}q_k$	b
	$\mathtt{ba}^{k-4}s_{k-1}$	b
	⋮	⋮
	$\mathtt{ba}s_4$	b
	$\mathtt{ba}s_3$	a
⋮	⋮	⋮
	$\mathtt{a}e_{k-1}$	a
$\mathtt{b}^{k-1}\mathtt{a}$	s_2	b
	$\mathtt{ba}^{k-3}q_k$	a
$\mathtt{b}^{k}\mathtt{a}$	s_2	a

The *block prefix* column shows the common prefix shared by all the rotations in a block. The *ordering factor* column shows the factor following the block prefix of a rotation, which decides its relative order inside its block. The BWT column shows the last character of each rotation. The dashed lines divide sub-ranges of rotations for which the BWT follows distinct patterns

Lemma 22 ($\beta(\mathtt{a}^{k-2}\mathtt{b})$) *Given the word* $w_k = (\prod_{i=2}^{k-1} s_i e_i)q_k$ *for some* $k > 5$, *the first rotation in the BWT matrix is* $\mathtt{a}^{k-3}q_k \cdots \mathtt{b}$.

Proof The first rotation in lexicographic order must start with the longest run of a's. By definition of w_k, the longest run of a's has length $k-2$, and it is obtained by concatenating the suffix $\mathtt{a}^{k-3}$ of e_{k-1} with q_k, which is preceded by a b (otherwise we could extend the run of a's). □

Lemma 23 ($\beta(\mathtt{a}^i\mathtt{b})$ *for* $4 \leq i \leq k-3$) *Given the word* $w_k = (\prod_{i=2}^{k-1} s_i e_i)q_k$ *for some* $k > 5$, *and an integer* $4 \leq i \leq k-3$, *the rotations in the BWT matrix starting with* $\mathtt{a}^i\mathtt{b}$ *are* $\mathtt{a}^{i-1}s_{i+2}\cdots\mathtt{b} < \mathtt{a}^{i-1}s_{i+3}\cdots\mathtt{a} < \ldots < \mathtt{a}^{i-1}s_{k-1}\cdots\mathtt{a} < \mathtt{a}^{i-1}q_k\cdots\mathtt{a}$.

Proof One can notice that, for all $4 \leq i \leq k-3$, the (circular) factor $\mathtt{a}^i\mathtt{b}$ can only be obtained, for all $i+2 \leq j \leq k$, from the concatenation of the suffix $\mathtt{a}^{i-1}$ of e_{j-1}, with either the prefix ab of s_j, if $i+2 \leq j \leq k-1$, or the prefix ab of q_k, if $j = k$. By Lemma 20, we can sort these rotations according to the lexicographic order of $\bigcup_{j=i}^{k-1}\{s_j\} \cup \{q_k\}$. Note that all these rotations end with an a, with the exception of the rotation starting with $\mathtt{a}^{i-1}s_{i+2}$, since it is where the only occurrence of $\mathtt{ba}^i\mathtt{b}$ can be found. □

Lemma 24 ($\beta(\mathtt{aaab})$) *Given the word* $w_k = (\prod_{i=2}^{k-1} s_i e_i)q_k$ *for some* $k > 5$, *the first five rotations in the BWT matrix starting with* aaab *are* $\mathtt{aa}e_2\cdots\mathtt{b} < \mathtt{aa}e_3\cdots\mathtt{b} < \mathtt{aa}e_4\cdots\mathtt{b} < \mathtt{aa}s_5\cdots\mathtt{b} < \mathtt{aa}e_5\cdots\mathtt{b}$, *while the remaining are* $\mathtt{aa}s_6\cdots\mathtt{a} < \mathtt{aa}e_6\cdots\mathtt{b} < \ldots < \mathtt{aa}s_{k-1}\cdots\mathtt{a} < \mathtt{aa}e_{k-1}\cdots\mathtt{b} < \mathtt{aa}q_k\cdots\mathtt{a}$. □

Proof Analogously to the proof of Lemma 23, some of the rotations starting with aaab can be obtained, for all $5 \leq j \leq k$, from the concatenation of the suffix aa of e_{j-1}, with either the prefix ab of s_j, if $5 \leq j \leq k-1$, or the prefix ab of q_k, if $j = k$. However, in this case we have more rotations starting with aaab, that are those rotations starting with the suffix aa of $s_{j'}$ concatenated with the prefix ab of $e_{j'}$, for all $2 \leq j' \leq k-1$. Thus, all the rotations starting with aaab are sorted according to the lexicographic order of the words in $\bigcup_{j=5}^{k-1}\{s_j\} \cup \bigcup_{j'=2}^{k-1}\{e_{j'}\} \cup \{q_k\}$. Note that all the rotations starting either with $\mathtt{aa}s_j$, for all $6 \leq j \leq k-1$, or with $\mathtt{aa}q_k$, end with a. On the other hand, the rotations starting either with $\mathtt{aa}s_5$ or with $\mathtt{aa}e_j$, for all $2 \leq j \leq k-1$, end with a b. □

Lemma 25 ($\beta(\mathtt{aab})$) *Given the word* $w_k = (\prod_{i=2}^{k-1} s_i e_i)q_k$ *for some* $k > 5$, *the first five rotations in the BWT matrix starting with* aab *are* $\mathtt{a}s_2\cdots\mathtt{b} < \mathtt{a}e_2\cdots\mathtt{a} < \mathtt{a}e_3\cdots\mathtt{a} < \mathtt{a}s_4\cdots\mathtt{b} < \mathtt{a}e_4\cdots\mathtt{a}$, *while the remaining are* $\mathtt{a}s_5\cdots\mathtt{a} < \mathtt{a}e_5\cdots\mathtt{a} < \ldots < \mathtt{a}s_{k-1}\cdots\mathtt{a} < \mathtt{a}e_{k-1}\cdots\mathtt{a} < \mathtt{a}q_k\cdots\mathtt{a}$.

Proof Each of the rotations starting with aaab from Lemma 24 induces a rotation starting with aab, obtained by shifting on the left one character a. It follows that all of these rotations end with an a. The other rotations starting with aab are the one obtained by concatenating the suffix a of e_3 and the prefix ab of s_4, and the one obtained by concatenating the suffix a of q_k and the prefix ab of s_2. Moreover, both

the rotations end with a b. The thesis follows by sorting the rotations according to the lexicographic order of the words in $\{s_2\} \cup \bigcup_{j=4}^{k-1}\{s_j\} \cup \bigcup_{j'=2}^{k-1}\{e_{j'}\} \cup \{q_k\}$. □

Lemma 26 (β(ab)) *Given the word* $w_k = (\prod_{i=2}^{k-1} s_i e_i)q_k$ *for some* $k > 5$, *the first* $k-2$ *rotations in the BWT matrix starting with* ab *are* $\mathtt{aba}^{k-3}q_k \cdots \mathtt{b} < \mathtt{aba}^{k-4}s_{k-1} \cdots \mathtt{b} < \ldots < \mathtt{ab}s_3 \cdots \mathtt{b}$, *the following four rotations are* $s_2 \cdots \mathtt{a} < e_2 \cdots \mathtt{a} < s_3 \cdots \mathtt{b} < e_3 \cdots \mathtt{a}$, *and the remaining are* $s_4 \cdots \mathtt{a} < e_4 \cdots \mathtt{a} < \ldots < s_{k-1} \cdots \mathtt{a} < e_{k-1} \cdots \mathtt{a} < q_k \cdots \mathtt{a}$.

Proof For any two distinct integers $i, i' \geq 0$, we have that $\mathtt{aba}^i\mathtt{b} < \mathtt{aba}^{i'}\mathtt{b}$ if and only if $i > i'$. Thus, the first rotation in lexicographic order starting with ab is the one that is followed by the longest run of a's. The smallest of these rotations can be found by concatenating the suffix $\mathtt{aba}^{k-3}$ of e_{k-1} with the prefix ab of q_k, followed by the suffix $\mathtt{aba}^{i-2}$ of e_{i-1} concatenated with the prefix ab of s_i, for all $3 \leq i \leq k-1$ taken in decreasing order. By construction of e_i, for all $3 \leq i \leq k-1$, these rotations must end with a b.

The remaining rotations starting with ab are exactly those rotations having as prefix either s_i or e_i, for all $2 \leq i \leq k-1$, or q_k. Note that all of these rotations are obtained by shifting on the left one character a from the rotations starting with aab from Lemma 25, with the exception of the one starting with s_3. It follows that the latter ends with a b, while all the other rotations with an a. □

Lemma 27 (β(ba)) *Given the word* $w_k = (\prod_{i=2}^{k-1} s_i e_i)q_k$ *for some* $k > 5$, *the first* $k-5$ *rotations in the BWT matrix starting with* ba *are* $\mathtt{ba}^{k-3}q_k \cdots \mathtt{a} < \mathtt{ba}^{k-4}s_{k-1} \cdots \mathtt{a} < \ldots < \mathtt{ba}^3 s_6 \cdots \mathtt{a}$, *followed by* $\mathtt{baa}e_2 \cdots \mathtt{b} < \mathtt{baa}e_3 \cdots \mathtt{b} < \mathtt{baa}e_4 \cdots \mathtt{b} < \mathtt{baa}s_5 \cdots \mathtt{a} < \mathtt{baa}e_5 \cdots \mathtt{b}$, *then by* $\mathtt{baa}e_6 \cdots \mathtt{b} < \mathtt{baa}e_7 \cdots \mathtt{b} < \ldots < \mathtt{baa}e_{k-1} \cdots \mathtt{b} < \mathtt{bas}_2 \cdots \mathtt{b} < \mathtt{bas}_4 \cdots \mathtt{a}$, *and finally by* $\mathtt{baba}^{k-3}q_k \cdots \mathtt{b} < \mathtt{baba}^{k-4}s_{k-1} \cdots \mathtt{b} < \ldots < \mathtt{bab}s_3 \cdots \mathtt{b} < \mathtt{b}s_3 \cdots \mathtt{a}$.

Proof One can notice that we have as many circular occurrences of ba as the number of maximal (circular) runs of b's in w_k. Then, for all $2 \leq i \leq k-1$, we have (i) one run of b's in s_i, and (ii) two runs in e_i, and (iii) one run in q_k.

For the case (i), we have one rotation starting with $\mathtt{baa}e_i$, for each $2 \leq i \leq k-1$. Since each run of b's within each word from $\bigcup_{i=2}^{k-1}\{s_i\}$ is of length at least 2, all rotations in (i) end with a b.

For the case (ii), for all $2 \leq i \leq k-1$, we can distinguish between two sub-cases, based on where ba starts: if either (ii.a) from the first run of b's in e_i, or (ii.b) from the second one. For the case (ii.a), we can see that these rotations are of the type $\mathtt{baba}^{i-2}s_{i+1}$, if $2 \leq i < k-2$, and $\mathtt{baba}^{k-3}q_k$. Analogously to the case (i), each rotations for case (ii.a) end with a b. Each rotation in (ii.b) is obtained by shifting two characters on the right each rotation in (ii.a). Therefore, all of these rotations end with an a and have prefixes of the type $\mathtt{ba}^{i-2}s_{i+1}$, if $2 \leq i < k-2$, or $\mathtt{ba}^{k-3}q_k$.

For the case (iii), the rotation starting with ba in q_k has $\mathtt{bas}_2$ as a prefix, and it is preceded by a b.

Observe that only for (ii.b) we have rotations starting with baaaa. Hence, the first rotation in lexicographic order is the one starting with $\mathtt{ba}^{k-3}q_k$, followed by those starting with $\mathtt{ba}^{k-4}s_{k-1} < \mathtt{ba}^{k-5}s_{k-2} < \ldots < \mathtt{baaa}s_6$.

Among the remaining rotations, those having prefix $\mathtt{baaa}$ either start with $\mathtt{baa}s_5$ from (ii.b), or $\mathtt{baa}e_i$ from (i), for all $2 \leq i \leq k-1$. Thus, by Lemma 20, we can sort them according to the order of the words in $\{s_5\} \cup \bigcup_{i=2}^{k-1}\{e_i\}$. Then, the remaining rotations with prefix $\mathtt{baa}$ are those starting with $\mathtt{ba}s_2$ from (iii), and $\mathtt{ba}s_4$ from (ii.b). Finally, let us focus on the rotations from case (ii.a). These rotations are sorted according to the length of the run of $\mathtt{a}$'s following the common prefix $\mathtt{bab}$, similarly to the sorting of the rotations from the case (ii.b). The last rotation left is the one starting with $\mathtt{b}s_3$ from case (ii.b). Since this rotation is greater than each word from case (ii.a), this is the greatest rotation of w_k starting with $\mathtt{ba}$ and the thesis follows.□

Lemma 28 ($\beta(\mathtt{b}^j\mathtt{a})$ for all $2 \leq j \leq k-1$) *Given the word $w_k = (\prod_{i=2}^{k-1} s_i e_i)q_k$ for some $k > 5$, and an integer $2 \leq i \leq k-2$, the first $k-i$ rotations in the BWT matrix starting with $\mathtt{b}^i\mathtt{a}$ are $\mathtt{b}^i\mathtt{aa}e_i \cdots \mathtt{a} < \mathtt{b}^i\mathtt{aa}e_{i+1} \cdots \mathtt{b} < \ldots < \mathtt{b}^i\mathtt{aa}e_{k-1} \cdots \mathtt{b} < \mathtt{b}^i\mathtt{a}s_2 \cdots \mathtt{b}$, followed by $\mathtt{b}^i\mathtt{aba}^{k-3}q_k \cdots \mathtt{b} < \mathtt{b}^i\mathtt{aba}^{k-4}s_{k-1} \cdots \mathtt{b} < \ldots < \mathtt{b}^i\mathtt{aba}^{i-1}s_{i+2} \cdots \mathtt{b} < \mathtt{b}^i\mathtt{aba}^{i-2}s_{i+1} \cdots \mathtt{a}$.*

Proof All runs of $\mathtt{b}$'s of length at least $2 \leq i \leq k-2$, either appear in (i) s_j or (ii) e_j, for all $i \leq j \leq k-1$, or in (iii) q_k. Let us consider the three cases separately. For all $i \leq j \leq k-1$, the rotation starting within s_j (i) has as prefix $\mathtt{b}^i\mathtt{aa}e_j$. For all $i \leq j \leq k-2$, the rotation starting within e_j (ii) has as prefix $\mathtt{b}^i\mathtt{aba}^{j-2}s_{j+1}$, and for $j = k-1$ we have the rotation with prefix $\mathtt{b}^i\mathtt{aba}^{k-3}q_k$. Finally, the rotation starting within q_k (iii) has as prefix $\mathtt{b}^i\mathtt{a}s_2$.

By construction, we can see that first we have all the rotations from case (i) sorted according to the lexicographic order of the words in $\bigcup_{j=i}^{k-1}\{e_i\}$ (Lemma 20), then we have the rotation from case (iii), and finally the rotation from case (ii), sorted according to the decreasing length of the run of $\mathtt{a}$'s following the common prefix $\mathtt{b}^i\mathtt{ab}$.

Moreover, note that only when the run of $\mathtt{b}$'s is of length exactly i the rotation end with an $\mathtt{a}$. Thus, the only for the rotations ending with an $\mathtt{a}$ are those starting within s_i and e_i, i.e. those with prefix $\mathtt{b}^i\mathtt{a}e_i$ and $\mathtt{b}^i\mathtt{aba}^{i-2}s_{i+1}$. □

Lemma 29 ($\beta(\mathtt{b}^k\mathtt{a})$) *Given the word $w_k = (\prod_{i=2}^{k-1} s_i e_i)q_k$ for some $k > 5$, the last four rotations of the BWT matrix are $\mathtt{b}^{k-1}\mathtt{aa}e_{k-1} \cdots \mathtt{a} < \mathtt{b}^{k-1}\mathtt{a}s_2 \cdots \mathtt{b} < \mathtt{b}^{k-1}\mathtt{aba}^{k-3}q_k \cdots \mathtt{a} < \mathtt{b}^k\mathtt{a}s_2 \cdots \mathtt{a}$.*

Proof Observe that the only rotations with prefix $\mathtt{b}^{k-1}\mathtt{a}$ either start within s_{k-1}, or q_k, or e_{k-1}. These rotations have prefix respectively $\mathtt{b}^{k-1}\mathtt{aa}e_{k-1}$, $\mathtt{b}^{k-1}\mathtt{a}s_2$, and $\mathtt{b}^{k-1}\mathtt{aba}^{k-3}q_k$. One can see that these rotations taken in this order are already sorted, and only the rotation starting within q_k ends with a $\mathtt{b}$, while the other two with an $\mathtt{a}$. Finally, the only occurrence of $\mathtt{b}^k$ is within q_k. Hence, the last rotation in lexicographic order starts with $\mathtt{b}^k\mathtt{a}s_2$, and since the run of $\mathtt{b}$'s is maximal it ends with an $\mathtt{a}$, and the thesis follows. □

The following proposition puts together the BWT computations carried out for all blocks of consecutive rows, highlighting which prefixes are shared.

Proposition 30 *Given an integer $k > 5$, let $w_k = (\prod_{i=2}^{k-1} s_i e_i) q_k$. Then,*

$$\begin{aligned}
\beta(\mathtt{a}^i\mathtt{b}) &= \mathtt{b}\mathtt{a}^{k-i-2} \text{ for all } 4 \le i \le k-2,\\
\beta(\mathtt{a}^3\mathtt{b}) &= \mathtt{b}^5(\mathtt{a}\mathtt{b})^{k-6}\mathtt{a},\\
\beta(\mathtt{a}^2\mathtt{b}) &= \mathtt{b}\mathtt{a}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-8},\\
\beta(\mathtt{a}\mathtt{b}) &= \mathtt{b}^{k-2}\mathtt{a}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-6},\\
\beta(\mathtt{b}\mathtt{a}) &= \mathtt{a}^{k-5}\mathtt{b}\mathtt{b}\mathtt{b}\mathtt{a}\mathtt{b}^{k-4}\mathtt{a}\mathtt{b}^{k-2}\mathtt{a},\\
\beta(\mathtt{b}^j\mathtt{a}) &= \mathtt{a}\mathtt{b}^{2k-2j-1}\mathtt{a} \text{ for all } 2 \le j \le k-1, \text{ and}\\
\beta(\mathtt{b}^k\mathtt{a}) &= \mathtt{a}.
\end{aligned}$$

Hence, the BWT of the w_k is $\mathrm{BWT}(w_k) = \prod_{i=2}^{k-1} \beta(\mathtt{a}^{k-i}\mathtt{b}) \cdot \prod_{i=1}^{k} \beta(\mathtt{b}^i\mathtt{a})$. *Moreover, $r(w_k) = 6k - 12$.*

Proof The words $\beta(\mathtt{a}^{k-2}\mathtt{b})$, $\beta(\mathtt{a}^i\mathtt{b})$ for all $4 \le i \le k-2$, $\beta(\mathtt{a}^3\mathtt{b})$, $\beta(\mathtt{a}^2\mathtt{b})$, $\beta(\mathtt{a}\mathtt{b})$, $\beta(\mathtt{b}\mathtt{a})$, $\beta(\mathtt{b}^j\mathtt{a})$ for all $2 \le j \le k-1$, and $\beta(\mathtt{b}^k\mathtt{a})$, are the concatenations of the last characters of the rotations from Lemma 22, Lemma 23, Lemma 24, Lemma 25, Lemma 26, Lemma 27, Lemma 28, and Lemma 29 respectively. Moreover, every rotation used to build $\beta(\mathtt{a}^i\mathtt{b})$ is smaller than each rotation used to build $\beta(\mathtt{a}^{i'}\mathtt{b})$, for every $1 \le i' < i \le k-2$. Symmetrically, every rotation used to build $\beta(\mathtt{b}^j\mathtt{a})$ is greater than each rotation used to build $\beta(\mathtt{b}^{j'}\mathtt{a})$, for every $1 \le j' < j \le k$. Since we have considered all the disjoint ranges of rotations of w_k based on their common prefix, the word $\prod_{i=2}^{k-1} \beta(\mathtt{a}^{k-i}\mathtt{b}) \cdot \prod_{i=1}^{k} \beta(\mathtt{b}^i\mathtt{a})$ is the BWT of w_k.

With the structure of $\mathrm{BWT}(w_k)$, we can easily derive its number of runs. The word $\prod_{i=2}^{k-4}(\beta(\mathtt{a}^{k-i}\mathtt{b}))$ has exactly $2(k-6)$ runs: we start with 2 runs from $\beta(\mathtt{a}^{k-2}\mathtt{b})\beta(\mathtt{a}^{k-3}\mathtt{b}) = \mathtt{b}\mathtt{b}\mathtt{a}$, and then, concatenating each other $\beta(\mathtt{a}^i\mathtt{b})$ up to $\beta(\mathtt{a}^4\mathtt{b})$ adds 2 new runs each. It is easy to see that $\beta(\mathtt{a}\mathtt{a}\mathtt{a}\mathtt{b})$, $\beta(\mathtt{a}\mathtt{a}\mathtt{b})$, and $\beta(\mathtt{a}\mathtt{b})$, have $2(k-5)$, 4, and 4 runs, respectively. Moreover, the boundaries between these words do not merge, nor with $\beta(\mathtt{a}^4\mathtt{b})$ in the case of $\beta(\mathtt{a}\mathtt{a}\mathtt{a}\mathtt{b})$. The word $\beta(\mathtt{b}\mathtt{a})$ has exactly 7 runs but it merges with $\beta(\mathtt{a}\mathtt{b})$ and $\beta(\mathtt{b}\mathtt{b}\mathtt{a})$, hence we only charge 5 runs to this word. The remaining part of the BWT, i.e., $\prod_{i=2}^{k}(\beta(\mathtt{b}^i\mathtt{a}))$, has $2(k-2)+1$ runs: we start with 3 runs from $\beta(\mathtt{b}\mathtt{b}\mathtt{a})$, and then, concatenating each other $\beta(\mathtt{b}^i\mathtt{a})$ up to $\beta(\mathtt{b}^{k-1}\mathtt{a})$ adds 2 new runs each. The word $\beta(\mathtt{b}^k\mathtt{a})$ does not add new runs, as it consists only of an $\mathtt{a}$ that merges with the previous one. Overall, we have $2(k-6) + 2(k-5) + 4 + 4 + 5 + 2(k-2) + 1 = 6k - 12$, and the claim holds. □

4.2 BWT of w_k After an Edit Operation

The following lemmas describe the BWT of w_k after some specific edit operations are applied. Instead of proving the whole structure of the BWT from the beginning, we compare how the edit operation changes either the relative order or the last character of the rotations of w_k. To do so, in this part we use the notation $\beta(v)$ and $\beta^\star(v)$ to

denote the BWT in correspondence of the rotations with prefix $v \in \Sigma^*$ of w_k and w'_k respectively, where w'_k is obtained after applying to w_k an specific edit operation. The number of runs in the BWT of w'_k can easily be derived by comparing its BWT to the BWT of w_k, for which we explicitly counted the number of runs, so we omit that part of the proofs. All the edit operations on w_k we show in this subsection increase the number $r(w_k)$ by a $\Theta(k)$ additive factor. To give an intuition, this increment comes mainly from the $\beta^\star(\mathtt{b}^j\mathtt{a})$ ranges for $2 \leq j \leq k-2$, because for each one of the corresponding ranges $\beta(\mathtt{b}^j\mathtt{a}) = \mathtt{a}\mathtt{b}^{2k-2j-1}\mathtt{a}$ in BWT(w_k), one of the b's is either moved to the top or the bottom of the range, in a consistent manner for each j (it depends on the edit operation if the b goes to the top or the bottom of the range, but it is the same behavior for all the ranges considered). Then, two ranges that originally agreed on their last and first character in w_k are now separated by a b, adding this way 2 new runs for each j.

Lemma 31 (BWT of $w_k\mathtt{a}$) *Given an integer $k > 5$, for $w_k\mathtt{a}$ it holds that*

$$\begin{aligned}
\beta^\star(\mathtt{a}^i\mathtt{b}) &= \mathtt{b}\mathtt{a}^{k-i-2} \text{ for all } 4 \leq i \leq k-2,\\
\beta^\star(\mathtt{a}^3\mathtt{b}) &= \mathtt{b}\mathtt{b}^5(\mathtt{a}\mathtt{b})^{k-6}\mathtt{a},\\
\beta^\star(\mathtt{a}^2\mathtt{b}) &= \mathtt{a}\mathtt{a}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-8},\\
\beta^\star(\mathtt{a}\mathtt{b}) &= \mathtt{b}^{k-2}\mathtt{a}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-6},\\
\beta^\star(\mathtt{b}\mathtt{a}) &= \mathtt{a}^{k-5}\mathtt{b}\mathtt{b}\mathtt{b}\mathtt{b}\mathtt{a}\mathtt{b}^{k-5}\mathtt{a}\mathtt{b}^{k-2}\mathtt{a},\\
\beta^\star(\mathtt{b}^j\mathtt{a}) &= \mathtt{b}\mathtt{a}\mathtt{b}^{2k-2j-2}\mathtt{a} \text{ for all } 2 \leq j \leq k-1 \text{ and}\\
\beta^\star(\mathtt{b}^k\mathtt{a}) &= \mathtt{a}.
\end{aligned}$$

Hence, $\mathrm{BWT}(w_k\mathtt{a}) = \prod_{i=2}^{k-1} \beta^\star(\mathtt{a}^{k-i}\mathtt{b}) \cdot \prod_{i=1}^{k} \beta^\star(\mathtt{b}^i\mathtt{a})$. *Moreover, it holds that* $r(w_k\mathtt{a}) = 8k - 20$.

Proof By Lemmas 22 and 23, we can see that appending an a after q_k does not affect the BWT in the range of rotations having $\mathtt{a}^i\mathtt{b}$ as a prefix, for all $4 \leq i \leq k-2$. Thus, $\beta^\star(\mathtt{a}^i\mathtt{b}) = \beta(\mathtt{a}^i\mathtt{b})$ for all $4 \leq i \leq k-2$.

The rotation starting with $\mathtt{a}\mathtt{a}s_2$, which is not a circular factor of w_k, ends with a b. By Lemma 24, we can see that such a rotation is the smallest one with prefix aaab in lexicographic order, while the other rotations maintain their relative order. Therefore, $\beta^\star(\mathtt{aaab}) = \mathtt{b} \cdot \beta(\mathtt{aaab})$.

By Lemma 25, the rotation with prefix $\mathtt{a}s_2$ is still the smallest rotation starting with aab, with the difference that in this case, it ends with the last a of q_k. It follows that $\beta^\star(\mathtt{aab})$ is obtained by replacing the first b of $\beta(\mathtt{aab})$ with an a.

Both the order and the last symbol of all the rotations having as prefix ab described in Lemma 26 is not affected by the insertion of the a, and therefore $\beta^\star(\mathtt{ab}) = \beta(\mathtt{ab})$.

Let us now consider all the rotations of w_k with prefix $\mathtt{b}^j\mathtt{a}s_2$, for all $1 \leq j \leq k$. One can notice that $w_k\mathtt{a}$ does not have any rotation starting with $\mathtt{b}^j\mathtt{a}s_2$, for all $1 \leq j \leq k$,

but instead, it has rotations starting with $\mathtt{b}^j\mathtt{a}\mathtt{a}s_2$. Thus, for all $1 \leq j \leq k-1$, to obtain $\beta^\star(\mathtt{b}^j\mathtt{a})$ from $\beta(\mathtt{b}^j\mathtt{a})$ we have to remove the $\mathtt{b}$ in correspondence of the rotations starting with $\mathtt{b}^j\mathtt{a}s_2$, and add a $\mathtt{b}$ in correspondence of the rotations $\mathtt{b}^j\mathtt{a}\mathtt{a}s_2$. By Lemmas 27, 28, and 29, such rotations are placed right before the rotation starting with $\mathtt{b}^j\mathtt{a}\mathtt{a}e_2$.

Finally, the last rotation has still the same prefix $\mathtt{b}^k\mathtt{a}$ and ends with an $\mathtt{a}$, and the thesis follows. □

Lemma 32 (BWT of $\widehat{w_k}$) *Given an integer $k > 5$, for $\widehat{w_k}$ it holds that*

$$\begin{aligned}
\beta^\star(\mathtt{a}^i\mathtt{b}) &= \mathtt{b}\mathtt{a}^{k-i-2} \text{ for all } 4 \leq i \leq k-2,\\
\beta^\star(\mathtt{a}^3\mathtt{b}) &= \mathtt{b}^5(\mathtt{a}\mathtt{b})^{k-6}\mathtt{a},\\
\beta^\star(\mathtt{a}^2\mathtt{b}) &= \mathtt{a}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-8},\\
\beta^\star(\mathtt{a}\mathtt{b}) &= \mathtt{b}^{k-2}\mathtt{b}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-6},\\
\beta^\star(\mathtt{b}\mathtt{a}) &= \mathtt{a}^{k-5}\mathtt{b}\mathtt{b}\mathtt{b}\mathtt{a}\mathtt{b}^{k-5}\mathtt{a}\mathtt{b}^{k-2}\mathtt{b}\mathtt{a},\\
\beta^\star(\mathtt{b}^j\mathtt{a}) &= \mathtt{a}\mathtt{b}^{2k-2j-2}\mathtt{a}\mathtt{b} \text{ for all } 2 \leq j \leq k-1 \text{ and}\\
\beta^\star(\mathtt{b}^k\mathtt{a}) &= \mathtt{a}.
\end{aligned}$$

Hence, $\mathrm{BWT}(\widehat{w_k}) = \prod_{i=2}^{k-1} \beta^\star(\mathtt{a}^{k-i}\mathtt{b}) \cdot \prod_{i=1}^{k} \beta^\star(\mathtt{b}^i\mathtt{a})$. *Moreover, it holds that* $r(\widehat{w_k}) = 8k-20$.

Proof Analogously to the previous Lemma, if we look in Lemmas 22, 23, and 24, at the structure of the BWT in correspondence of the rotations starting with $\mathtt{a}^i\mathtt{b}$, for all $3 \leq i \leq k-2$, we can notice that the order or the symbols in the BWT is not affected. Thus, for all $3 \leq i \leq k-2$, we have $\beta^\star(\mathtt{a}^i\mathtt{b}) = \beta(\mathtt{a}^i\mathtt{b})$.

Since the last $\mathtt{a}$ of q_k is omitted, the circular factor $\mathtt{a}s_2$ does not appear anymore in $\widehat{w}$. Thus, $\beta^\star(\mathtt{a}\mathtt{a}\mathtt{b})$ is obtained by removing the first $\mathtt{b}$ from $\beta(\mathtt{a}\mathtt{a}\mathtt{b})$, since by Lemma 25 it is in correspondence of the rotation with prefix $\mathtt{a}s_2$.

On the other hand, we can observe from Lemma 26 that the rotation with prefix s_2 maintains its relative order also in $\widehat{w_k}$, but its last symbol is now a $\mathtt{b}$ instead of an $\mathtt{a}$.

For each $1 \leq j \leq k$, the rotation starting with $\mathtt{b}^j\mathtt{a}s_2$ of w_k does not appear in $\widehat{w_k}$, but in fact it is replaced by one having $\mathtt{b}^j s_2$ as prefix and ending in the same way. When $j = 1$, by Lemma 27 such a rotation is located between the last two rotations with the prefix $\mathtt{b}\mathtt{a}$, which start by $\mathtt{b}\mathtt{a}\mathtt{b}s_3$ and $\mathtt{b}s_3$ respectively. When $2 \leq j \leq k-1$, by Lemmas 28 and 29, the rotation starting with $\mathtt{b}^j s_2$ is greater than all the other rotations with prefix $\mathtt{b}^j\mathtt{a}$. Thus, for all $1 \leq j \leq k-1$, we obtain $\beta^\star(\mathtt{b}^j\mathtt{a})$ by moving the $\mathtt{b}$ in correspondence of the rotation starting with $\mathtt{b}\mathtt{a}s_2$ from $\beta(\mathtt{b}^j\mathtt{a})$ and placing it in correspondence of $\mathtt{b}^j s_2$. Finally, the last rotation has still the same prefix $\mathtt{b}^k\mathtt{a}$ and ends with an $\mathtt{a}$, and the thesis follows. □

Lemma 33 (BWT of $\widehat{w_k}\mathtt{b}$) *Given an integer $k > 5$, for $\widehat{w_k}\mathtt{b}$ it holds that*

$$\begin{aligned}
\beta^\star(\mathtt{a}^i\mathtt{b}) &= \mathtt{b}\mathtt{a}^{k-i-2} \text{ for all } 4 \le i \le k-2,\\
\beta^\star(\mathtt{a}^3\mathtt{b}) &= \mathtt{b}^5(\mathtt{a}\mathtt{b})^{k-6}\mathtt{a},\\
\beta^\star(\mathtt{a}^2\mathtt{b}) &= \mathtt{a}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-8},\\
\beta^\star(\mathtt{a}\mathtt{b}) &= \mathtt{b}^{k-2}\mathtt{b}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-6},\\
\beta^\star(\mathtt{b}\mathtt{a}) &= \mathtt{a}^{k-5}\mathtt{b}\mathtt{b}\mathtt{b}\mathtt{a}\mathtt{b}^{k-5}\mathtt{a}\mathtt{b}^{k-2}\mathtt{b}\mathtt{a},\\
\beta^\star(\mathtt{b}^j\mathtt{a}) &= \mathtt{a}\mathtt{b}^{2k-2j-2}\mathtt{a}\mathtt{b} \text{ for all } 2 \le j \le k-1,\\
\beta^\star(\mathtt{b}^k\mathtt{a}) &= \mathtt{b} \text{ and}\\
\beta^\star(\mathtt{b}^{k+1}\mathtt{a}) &= \mathtt{a}.
\end{aligned}$$

Hence, $\mathrm{BWT}(\widehat{w_k}\mathtt{b}) = \prod_{i=2}^{k-1}\beta^\star(\mathtt{a}^{k-i}\mathtt{b}) \cdot \prod_{i=1}^{k+1}\beta^\star(\mathtt{b}^i\mathtt{a})$. *Moreover, it holds that* $r(\widehat{w_k}\mathtt{b}) = 8k - 20$.

Proof For the rotations in correspondence of the rotations starting with an $\mathtt{a}$, notice that replacing the last $\mathtt{a}$ of w_k for a $\mathtt{b}$ or removing the last $\mathtt{a}$ affects the BWT in the same way. Therefore, $\beta^\star(\mathtt{a}^i\mathtt{b})$ is the same as Lemma 32 for all $1 \le i \le k-2$.

The same behaviour can be noticed on the rotations with prefix $\mathtt{b}^j\mathtt{a}$, for all $1 \le j \le k-1$, while the rotation starting with $\mathtt{b}^k\mathtt{a}$ is now preceded by a $\mathtt{b}$.

With respect to the other edit operations, we have the range of rotations starting with $\mathtt{b}^{k+1}\mathtt{a}$, which consists solely in $\mathtt{b}^{k+1}s_2 \cdots \mathtt{a}$. □

The structure of the BWT of w_k and other words obtained by applying one or more edit operations on w_k are summed up in Table 2.

For a given word $w \neq \epsilon$, let w^{ins}, w^{del}, and w^{sub} be the words obtained by applying on w an insertion, a deletion, and a substitution of a character respectively.

We compare the number of runs of w_k and its variations and notice that the difference after applying one of the edit operations is $\Theta(k)$ in the three cases.

Proposition 34 *There exists an infinite family of words w such that: (i) $r(w^{ins}) - r(w) = \Theta(\sqrt{n})$; (ii) $r(w^{del}) - r(w) = \Theta(\sqrt{n})$; (iii) $r(w^{sub}) - r(w) = \Theta(\sqrt{n})$.*

Proof The family is composed of the words w_k with $k > 5$. Let $n = |w_k|$. If $w_k^{ins} = w_k\mathtt{a}$, $w_k^{del} = \widehat{w_k}$, and $w_k^{sub} = \widehat{w_k}\mathtt{b}$, from Proposition 30, Lemma 31, Lemma 32, and Lemma 33, we have that $r(w_k\mathtt{a}) = r(\widehat{w_k}) = r(\widehat{w_k}\mathtt{b}) = r(w_k) + (2k - 8)$. From Observation 2, we have that $2k - 8 = \Theta(\sqrt{n})$. □

5 Bit Catastrophes for $r_\$$

In this section, we discuss bit catastrophes when the parameter $r_\$$ is considered. Recall that for a word v, $r_\$(v) = runs(\mathrm{BWT}(v\$))$.

Table 2 BWTs of the word w_k and its variants after different edit operations

Word	$\beta(\$)$	$\beta(a\$)$	$\beta(aa\$)$	$\beta(a^ib)$	$\beta(a^3b)$	$\beta(a^2b)$	$\beta(ab)$
w_k	ϵ	ϵ	ϵ	ba^{k-i-2}	$b^5(ab)^{k-6}a$	$baaba^{2k-8}$	$b^{k-2}aaba^{2k-6}$
w_ka	ϵ	ϵ	ϵ	ba^{k-i-2}	$bb^5(ab)^{k-6}a$	$aaaba^{2k-8}$	$b^{k-2}aaba^{2k-6}$
$\widehat{w_k}$	ϵ	ϵ	ϵ	ba^{k-i-2}	$b^5(ab)^{k-6}a$	$aaba^{2k-8}$	$b^{k-2}baba^{2k-6}$
$\widehat{w_k}b$	ϵ	ϵ	ϵ	ba^{k-i-2}	$b^5(ab)^{k-6}a$	$aaba^{2k-8}$	$b^{k-2}baba^{2k-6}$
$w_k\$$	a	b	ϵ	ba^{k-i-2}	$b^5(ab)^{k-6}a$	$aaba^{2k-8}$	$b^{k-2}\$aba^{2k-6}$
$w_kb\$$	b	ϵ	ϵ	ba^{k-i-2}	$b^5(ab)^{k-6}a$	$aaba^{2k-8}$	$bb^{k-2}\$aba^{2k-6}$
$w_kbb\$$	b	ϵ	ϵ	ba^{k-i-2}	$b^5(ab)^{k-6}a$	$aaba^{2k-8}$	$bb^{k-2}\$aba^{2k-6}$
$w_ka\$$	a	a	b	ba^{k-i-2}	$b^5(ab)^{k-6}a$	$aaba^{2k-8}$	$b^{k-2}\$aba^{2k-6}$
Word	$\beta(b\$)$	$\beta(ba)$	$\beta(bb\$)$	$\beta(b^ja)$	$\beta(b^ka)$	$\beta(b^{k+1})$	$r(\cdot)$
w_k	ϵ	$a^{k-5}bbbab^{k-4}ab^{k-2}a$	ϵ	$ab^{2k-2j-1}a$	a	ϵ	$6k-12$
w_ka	ϵ	$a^{k-5}bbbbab^{k-5}ab^{k-2}a$	ϵ	$bab^{2k-2j-2}a$	a	ϵ	$8k-20$
$\widehat{w_k}$	ϵ	$a^{k-5}bbbab^{k-5}ab^{k-2}ba$	ϵ	$ab^{2k-2j-2}ab$	a	ϵ	$8k-20$
$\widehat{w_k}b$	ϵ	$a^{k-5}bbbab^{k-5}ab^{k-2}ba$	ϵ	$ab^{2k-2j-2}ab$	b	a	$8k-20$
$w_k\$$	ϵ	$ba^{k-5}bbbab^{k-5}ab^{k-2}a$	ϵ	$bab^{2k-2j-2}a$	a	ϵ	$8k-16$
$w_kb\$$	a	$a^{k-5}bbbab^{k-5}abb^{k-2}a$	ϵ	$ab^{2k-2j-1}a$	a	ϵ	$6k-13$
$w_kbb\$$	b	$a^{k-5}bbbab^{k-5}abb^{k-2}a$	a	$ab^{2k-2j-2}ab$	a	ϵ	$8k-17$
$w_ka\$$	ϵ	$ba^{k-5}bbbab^{k-5}ab^{k-2}a$	ϵ	$bab^{2k-2j-2}a$	a	ϵ	$8k-16$

The word in the intersection of the column $\beta(x)$ with the row w is the range of BWT(w) corresponding to all the rotations that have x as a prefix. The columns $\beta(a^ib)$ and $\beta(b^ja)$ represent ranges of columns from $i \in [k-2, 4]$ (in that order) and $j \in [2, k-1]$, respectively. Note that the prefixes in the columns are disjoint, and cover all the possible ranges for the set of words considered. The BWT of each word is the concatenation of all the words in its row from left to right. In the last column appears the number of BWT-runs of each of these words

5.1 When There is No Bit Catastrophe for $r_\$$

First, let us consider the case where a symbol $c \in \Sigma$ is prepended to a word v. As recently noted in [1], it is well known that in this case the value $r_\$$ can only vary by a constant value. For the sake of completeness, we include a proof.

Proposition 35 *For any $c \in \Sigma$, we have $r_\$(v) - 1 \leq r_\$(cv) \leq r_\$(v) + 2$.*

Proof Let us consider the list of lexicographically sorted cyclic rotations or, equivalently, the list of lexicographically sorted suffixes of $cv\$$. (The equivalence follows from the fact that \$ is smaller than all other characters.) This list can be obtained from the list of suffixes of $v\$$, to which the suffix $cv\$$ is added. Note that the relative order of all suffixes other than $cv\$$ remains the same. Moreover, the corresponding symbols in the BWT also remain the same, except that the character c takes the place of \$. This replacement decreases the number of BWT-runs by 0, 1, or 2, depending on whether this position in the BWT is preceded by a run of c, followed by a run of c, or both. The symbol corresponding to the new suffix $cv\$$ (which produces the insertion of \$ in the corresponding position in the BWT) increases the number of BWT-runs by 1 (if it is inserted between two existing runs), or by 2 (if it breaks a run). □

The following proposition shows that there are some cases in which $r_\$$ is not affected by any bit catastrophe.

Proposition 36 *Let c be smaller than or equal to the smallest character in a word v, then $r_\$(v) \leq r_\$(vc) \leq r_\$(v) + 1$.*

Proof The rotations of $vc\$$ can be viewed as the rotation $\$vc$, plus the rotations of $v\$$, where the occurrence of \$ has been replaced by $c\$$. The smallest of these is of course $\$vc$, since it starts with \$, while all others appear in the same order as before. This is because c is smaller or equal to the smallest character of v and greater than \$, and therefore, replacing \$ by $c\$$ does not change the lexicographic order of these rotations. This implies $\mathrm{BWT}(vc\$) = c \cdot \mathrm{BWT}(v\$)$, and thus, $r_\$(v) \leq r_\$(vc) \leq r_\$(v) + 1$. □

5.2 Multiplicative Bit Catastrophes for $r_\$$

We can derive from our results in Sec. 3 that there exist families of strings on which an edit operation can result in an increase of $r_\$$ by a multiplicative factor of $\log n$.

Proposition 37 *Let v be the Lyndon rotation of the Fibonacci word s of even order $2k > 4$, and $n = |v|$. Let v' be the word resulting by appending a* `b` *to v. Then $r_\$(v') = \Theta(\log n)$.*

Proof Let $s = x_{2k}\mathtt{ab} = x_{2k-1}\mathtt{b}\mathtt{a}x_{2k-2}\mathtt{ab}$ be the Fibonacci word of order $2k$. One can see that $v = \mathtt{a}x_{2k}\mathtt{b} = \mathtt{a}x_{2k-2}\mathtt{ab}x_{2k-1}\mathtt{b}$ [4]. Since v is a rotation of s, it holds that $r(v) = 2$. By using Lemma 16, $r_\$(v) = \Theta(1)$ since v is a Lyndon word. When we append `b` to v, we obtain $v' = \mathtt{a}x_{2k-2}\mathtt{ab}x_{2k-1}\mathtt{bb}$. One can note that $v' = \mathtt{a}x_{2k-2}\mathtt{ab}x_{2k-1}\mathtt{bb}$ is also a Lyndon word. Moreover, appending `b` to v is equivalent to inserting `b` in s at position $F_{2k-1} - 2$, implying that v' is a rotation of s',

where s' is s with a $\mathtt{b}$ inserted in position $F_{2k-1} - 2$. By using Proposition 3, we thus have that $r(v') = r(s') = \Theta(\log n)$. Since v' is also a Lyndon word, therefore $r_{\$}(v') = \Theta(\log n)$, using Lemma 16 again. □

5.3 Additive Bit Catastrophes for $r_{\$}$

In general, appending, deleting, or substituting with a symbol that is not the smallest of the alphabet can increase the number of runs of a word by an additive factor of $\Theta(\sqrt{n})$.

Lemma 38 (BWT of $w_k\$$) *Given an integer $k > 5$, for $w_k\$$ it holds that*

$$\begin{aligned}
\beta^\star(\$) &= \mathtt{a}\\
\beta^\star(\mathtt{a}\$) &= \mathtt{b}\\
\beta^\star(\mathtt{a}^i\mathtt{b}) &= \mathtt{b}\mathtt{a}^{k-i-2} \textit{ for all } 4 \le i \le k-2,\\
\beta^\star(\mathtt{a}^3\mathtt{b}) &= \mathtt{b}^5(\mathtt{a}\mathtt{b})^{k-6}\mathtt{a},\\
\beta^\star(\mathtt{a}^2\mathtt{b}) &= \mathtt{a}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-8},\\
\beta^\star(\mathtt{a}\mathtt{b}) &= \mathtt{b}^{k-2}\$\mathtt{a}\mathtt{b}\mathtt{a}^{2k-6},\\
\beta^\star(\mathtt{b}\mathtt{a}) &= \mathtt{b}\mathtt{a}^{k-5}\mathtt{b}\mathtt{b}\mathtt{b}\mathtt{a}\mathtt{b}^{k-5}\mathtt{a}\mathtt{b}^{k-2}\mathtt{a},\\
\beta^\star(\mathtt{b}^j\mathtt{a}) &= \mathtt{b}\mathtt{a}\mathtt{b}^{2k-2j-2}\mathtt{a} \textit{ for all } 2 \le j \le k-1 \textit{ and}\\
\beta^\star(\mathtt{b}^k\mathtt{a}) &= \mathtt{a}.
\end{aligned}$$

Hence, $\mathrm{BWT}(w_k\$) = \beta^\star(\$) \cdot \beta^\star(\mathtt{a}\$) \cdot \prod_{i=2}^{k-1} \beta^\star(\mathtt{a}^{k-i}\mathtt{b}) \cdot \prod_{i=1}^{k} \beta^\star(\mathtt{b}^i\mathtt{a})$. *Moreover, it holds that* $r(w_k\$) = 8k - 16$.

Proof The first rotation of $\mathrm{BWT}(w_k\$)$ is $\$w_k$ and ends with an $\mathtt{a}$ because w_k ends with an $\mathtt{a}$. Hence, $\beta^\star(\$) = \mathtt{a}$. There is also a rotation $\mathtt{a}\$\widehat{w_k}$, which ends with a $\mathtt{b}$ because $\widehat{w_k}$ ends with a $\mathtt{b}$. Hence, $\beta^\star(\mathtt{a}\$) = \mathtt{b}$. It is left to compare the remaining ranges $\beta^\star(v)$ with respect to $\beta(v)$.

It is easy to see from Lemma 22, Lemma 23, and Lemma 24 that $\beta^\star(\mathtt{a}^i\mathtt{b}) = \beta(\mathtt{a}^i\mathtt{b})$ for all $3 \le i \le k-2$.

The rotation starting with $\mathtt{a}s_2$ in w_k does not exist anymore when \$ is appended to w_k. By Lemma 25 the remaining rotations keep their last symbols and relative order. Therefore, $\beta^\star(\mathtt{a}\mathtt{a}\mathtt{b})$ is the same as $\beta(\mathtt{a}\mathtt{a}\mathtt{b})$ but with the first character removed, i.e., $\beta^\star(\mathtt{a}\mathtt{a}\mathtt{b}) = \mathtt{a}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-8}$.

For the rotations starting with $\mathtt{a}\mathtt{b}$, it happens that the rotation that originally started with s_2 in w_k, now ends with a \$. By Lemma 26, the remaining rotations do not change their last symbol. Also, all the rotations keep their relative order. Hence, $\beta^\star(\mathtt{a}\mathtt{b}) = \mathtt{b}^{k-2}\$\mathtt{a}\mathtt{b}\mathtt{a}^{2k-6}$.

In the case of the rotations starting with $\mathtt{b}\mathtt{a}$, the rotation that originally started with $\mathtt{b}\mathtt{a}s_2$ now starts with $\mathtt{b}\mathtt{a}\$s_2$ and is the smallest of its range. From Lemma 27 the remaining rotations keep their last symbols and relative order. Hence, $\beta^\star(\mathtt{b}\mathtt{a}) = \mathtt{b}\mathtt{a}^{k-5}\mathtt{b}\mathtt{b}\mathtt{b}\mathtt{a}\mathtt{b}^{k-5}\mathtt{a}\mathtt{b}^{k-2}\mathtt{a}$.

For the rotations starting with $\mathtt{b}^j\mathtt{a}$ for $2 \leq j \leq k-1$, one can notice that after appending \$ to w_k, the rotation that previously started with $\mathtt{b}^j\mathtt{a}s_2$ and ended with a b, now starts with $\mathtt{b}^j\mathtt{a}\$s_2$ and still ends with a b. Moreover, this rotation is smaller than the rotation starting with $\mathtt{b}^j\mathtt{a}\mathtt{a}e_j$. From Lemma 28 and Lemma 29 we can see that all the other rotations keep their relative order and last symbols. The rotation starting with $\mathtt{b}^j\mathtt{a}\mathtt{a}e_j$ still ends with an a, but now is the second smallest of its range. Hence, $\beta^\star(\mathtt{b}^j\mathtt{a}) = \mathtt{b}\mathtt{a}\mathtt{b}^{2k-2j-2}\mathtt{a}$ for all $2 \leq j \leq k-1$.

Finally, it is clear that $\beta^\star(\mathtt{b}^k\mathtt{a}) = \mathtt{a}$, as there is only one maximal run of k symbol b's, and it is not preceded by \$. □

Lemma 39 (BWT of $w_k\mathtt{b}\$$) *Given an integer $k > 5$, for $w_k\mathtt{b}\$$ it holds that*

$$\begin{aligned}
\beta^\star(\$) &= \mathtt{b}\\
\beta^\star(\mathtt{a}^i\mathtt{b}) &= \mathtt{b}\mathtt{a}^{k-i-2} \text{ for all } 4 \leq i \leq k-2,\\
\beta^\star(\mathtt{a}^3\mathtt{b}) &= \mathtt{b}^5(\mathtt{a}\mathtt{b})^{k-6}\mathtt{a},\\
\beta^\star(\mathtt{a}^2\mathtt{b}) &= \mathtt{a}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-8},\\
\beta^\star(\mathtt{a}\mathtt{b}) &= \mathtt{b}\mathtt{b}^{k-2}\$\mathtt{a}\mathtt{b}\mathtt{a}^{2k-6},\\
\beta^\star(\mathtt{b}\$) &= \mathtt{a},\\
\beta^\star(\mathtt{b}\mathtt{a}) &= \mathtt{a}^{k-5}\mathtt{b}\mathtt{b}\mathtt{b}\mathtt{a}\mathtt{b}^{k-5}\mathtt{a}\mathtt{b}\mathtt{b}^{k-2}\mathtt{a},\\
\beta^\star(\mathtt{b}^j\mathtt{a}) &= \mathtt{a}\mathtt{b}^{2k-2j-1}\mathtt{a} \text{ for all } 2 \leq j \leq k-1 \text{ and}\\
\beta^\star(\mathtt{b}^k\mathtt{a}) &= \mathtt{a}.
\end{aligned}$$

Hence, $\mathrm{BWT}(w_k\mathtt{b}\$) = \beta^\star(\$)\cdot(\prod_{i=2}^{k-1}\beta^\star(\mathtt{a}^{k-i}\mathtt{b}))\cdot\beta^\star(\mathtt{b}\$)\cdot(\prod_{i=1}^{k}\beta^\star(\mathtt{b}^i\mathtt{a}))$. *Moreover, it holds that* $r(w_k\mathtt{b}\$) = 6k-13$.

Proof The first rotation of $\mathrm{BWT}(w_k\mathtt{b}\$)$ is $\$w_k\mathtt{b}$. Hence, $\beta^\star(\$) = \mathtt{b}$. There is also a rotation $\mathtt{b}\$w_k$, which ends with an a because w_k ends with an a. Hence, $\beta^\star(\mathtt{b}\$) = \mathtt{a}$. It is left to compare the remaining ranges $\beta^\star(v)$ with respect to $\beta(v)$.

It is easy to see from Lemma 22, Lemma 23, and Lemma 24 that $\beta^\star(\mathtt{a}^i\mathtt{b}) = \beta(\mathtt{a}^i\mathtt{b})$ for all $3 \leq i \leq k-2$.

The rotation starting with $\mathtt{a}s_2$ in w_k does not exist anymore when b\$ is appended to w_k. By Lemma 25 the remaining rotations keep their last symbols and relative order. Therefore, $\beta^\star(\mathtt{aab})$ is the same as $\beta(\mathtt{aab})$ but with the first character removed, i.e., $\beta^\star(\mathtt{aab}) = \mathtt{a}\mathtt{a}\mathtt{b}\mathtt{a}^{2k-8}$.

For the rotations starting with ab, it happens that the rotation that originally started with s_2 in w_k, now ends with a \$ when b\$ is appended. Also, there is a new rotation starting with ab\$ that ends with b, and is clearly the smallest of the range. By Lemma 26, the remaining rotations do not change their last symbol. Also, all the rotations that come from w_k keep their relative order. Hence, $\beta^\star(\mathtt{ab}) = \mathtt{b}\mathtt{b}^{k-2}\$\mathtt{a}\mathtt{b}\mathtt{a}^{2k-6}$.

In the case of the rotations starting with ba, the rotation that originally started with $\mathtt{ba}s_2$ now starts with $\mathtt{bab}\$s_2$ and can be found just before the rotation starting with $\mathtt{baba}^{k-2}$. From Lemma 27 the remaining rotations keep their last symbols and relative order. Hence, $\beta^\star(\mathtt{ba}) = \mathtt{a}^{k-5}\mathtt{b}\mathtt{b}\mathtt{b}\mathtt{a}\mathtt{b}^{k-5}\mathtt{a}\mathtt{b}\mathtt{b}^{k-2}\mathtt{a}$.

For the rotations starting with $\mathtt{b}^j\mathtt{a}$ for $2 \le j \le k-1$, one can notice that after appending $\mathtt{b\$}$ to w_k, the rotation that previously started with $\mathtt{b}^j\mathtt{a}s_2$ and ended with a $\mathtt{b}$, now starts with $\mathtt{b}^j\mathtt{ab\$}s_2$ and still ends with a $\mathtt{b}$. Moreover, this rotation is still strictly in between the rotations starting with $\mathtt{b}^j\mathtt{aa}e_j$ and $\mathtt{b}^j\mathtt{aba}^{j-2}s_{j+1}$ (q_k instead of s_{j+1} if $j = k-1$). From Lemma 28 and Lemma 29, we can see that the latter two rotations are still the smallest and greatest of the range, and both end with an $\mathtt{a}$. Also, all the other rotations keep their last symbols. Hence, $\beta^\star(\mathtt{b}^j\mathtt{a}) = \beta(\mathtt{b}^j\mathtt{a})$ for all $2 \le j \le k-1$.

Finally, it is clear that $\beta^\star(\mathtt{b}^k\mathtt{a}) = \mathtt{a}$, as there is only one maximal run of k symbol $\mathtt{b}$'s, and it is not preceded by \$. □

Lemma 40 (BWT of $w_k\mathtt{bb\$}$) *Given an integer $k > 5$, for $w_k\mathtt{bb\$}$ it holds that*

$$\begin{aligned}
\beta^\star(\$) &= \mathtt{b}\\
\beta^\star(\mathtt{a}^i\mathtt{b}) &= \mathtt{ba}^{k-i-2} \textit{ for all } 4 \le i \le k-2,\\
\beta^\star(\mathtt{a}^3\mathtt{b}) &= \mathtt{b}^5(\mathtt{ab})^{k-6}\mathtt{a},\\
\beta^\star(\mathtt{a}^2\mathtt{b}) &= \mathtt{aaba}^{2k-8},\\
\beta^\star(\mathtt{ab}) &= \mathtt{bb}^{k-2}\mathtt{\$aba}^{2k-6},\\
\beta^\star(\mathtt{b\$}) &= \mathtt{b},\\
\beta^\star(\mathtt{ba}) &= \mathtt{a}^{k-5}\mathtt{bbbab}^{k-5}\mathtt{abb}^{k-2}\mathtt{a},\\
\beta^\star(\mathtt{bb\$}) &= \mathtt{a},\\
\beta^\star(\mathtt{b}^j\mathtt{a}) &= \mathtt{ab}^{2k-2j-2}\mathtt{ab} \textit{ for all } 2 \le j \le k-1 \textit{ and}\\
\beta^\star(\mathtt{b}^k\mathtt{a}) &= \mathtt{a}.
\end{aligned}$$

Hence, $\mathrm{BWT}(w_k\mathtt{bb\$}) = \beta^\star(\$) \cdot (\prod_{i=2}^{k-1} \beta^\star(\mathtt{a}^{k-i}\mathtt{b})) \cdot \beta^\star(\mathtt{b\$}) \cdot \beta^\star(\mathtt{ba}) \cdot \beta^\star(\mathtt{bb\$}) \cdot (\prod_{i=2}^{k} \beta^\star(\mathtt{b}^i\mathtt{a}))$. *Moreover, it holds that* $r(w_k\mathtt{b\$}) = 8k-17$.

Proof The first rotation of $\mathrm{BWT}(w_k\mathtt{bb\$})$ is $\$w_k\mathtt{bb}$. Hence, $\beta^\star(\$) = \mathtt{b}$. There is another new rotation $\mathtt{b}\$w_k\mathtt{b}$. Hence, $\beta^\star(\mathtt{b\$}) = \mathtt{b}$. There is also a rotation $\mathtt{bb}\$w_k$ that ends with an $\mathtt{a}$ because w_k ends with an $\mathtt{a}$. Hence, $\beta^\star(\mathtt{bb\$}) = \mathtt{a}$. It is left to compare the remaining ranges $\beta^\star(v)$ with respect to $\beta(v)$.

It is easy to see from Lemma 22, Lemma 23, and Lemma 24 that $\beta^\star(\mathtt{a}^i\mathtt{b}) = \beta(\mathtt{a}^i\mathtt{b})$ for all $3 \le i \le k-2$.

The rotation starting with $\mathtt{a}s_2$ in w_k does not exist anymore when $\mathtt{bb\$}$ is appended to w_k. By Lemma 25 the remaining rotations keep their last symbols and relative order. Therefore, $\beta^\star(\mathtt{aab})$ is the same as $\beta(\mathtt{aab})$ but with the first character removed, i.e., $\beta^\star(\mathtt{aab}) = \mathtt{aaba}^{2k-8}$.

For the rotations starting with $\mathtt{ab}$, it happens that the rotation that originally started with s_2 in w_k, now ends with a \$ when $\mathtt{bb\$}$ is appended. Also, there is a new rotation starting with $\mathtt{abb\$}$ that ends with $\mathtt{b}$, and can be found just before the rotation starting with s_2. By Lemma 26, the remaining rotations do not change their last symbol. Also, all the rotations that come from w_k keep their relative order. Hence, $\beta^\star(\mathtt{ab}) = \mathtt{b}^{k-2}\mathtt{b\$aba}^{2k-6}$.

In the case of the rotations starting with $\mathtt{ba}$, the rotation that originally started with $\mathtt{ba}s_2$ now starts with $\mathtt{babb\$}s_2$ and can be found just before the rotation starting with $\mathtt{b}s_3$ (the greatest on the range). From Lemma 27 we can see that the remaining rotations keep their last symbols and relative order. Hence, $\beta^\star(\mathtt{ba}) = \mathtt{a}^{k-5}\mathtt{bbbab}^{k-5}\mathtt{ab}^{k-2}\mathtt{ba}$.

For the rotations starting with $\mathtt{b}^j\mathtt{a}$ for $2 \leq j \leq k-1$, one can notice that after appending $\mathtt{bb\$}$ to w_k, the rotation that previously started with $\mathtt{b}^j\mathtt{a}s_2$ and ended with a $\mathtt{b}$, now starts with $\mathtt{b}^j\mathtt{abb\$}s_2$ and still ends with a $\mathtt{b}$. Moreover, this rotation is greater than the rotation starting with $\mathtt{b}^j\mathtt{aba}^{j-2}s_{j+1}$ (q_k instead of s_{j+1} if $j = k-1$). From Lemma 28 and Lemma 29 we can see that all the other rotations keep their relative order an last symbols. The rotation starting with $\mathtt{b}^j\mathtt{aba}^{j-2}s_{j+1}$ (q_k instead of s_{j+1} if $j = k-1$) still ends with an $\mathtt{a}$, but now is the second greatest of its range. Hence, $\beta^\star(\mathtt{b}^j\mathtt{a}) = \mathtt{ab}^{2k-2j-2}\mathtt{ab}$ for all $2 \leq j \leq k-1$.

Finally, it is clear that $\beta^\star(\mathtt{b}^k\mathtt{a}) = \mathtt{a}$, as there is only one maximal run of k symbol $\mathtt{b}$'s, and it is not preceded by \$. □

Lemma 41 (BWT of $w_k\mathtt{a\$}$) *Given an integer $k > 5$, for $w_k\mathtt{a\$}$ it holds that*

$$\begin{aligned}
\beta^\star(\mathtt{\$}) &= \mathtt{a}\\
\beta^\star(\mathtt{a\$}) &= \mathtt{a}\\
\beta^\star(\mathtt{aa\$}) &= \mathtt{b}\\
\beta^\star(\mathtt{a}^i\mathtt{b}) &= \mathtt{ba}^{k-i-2} \textit{ for all } 4 \leq i \leq k-2,\\
\beta^\star(\mathtt{a}^3\mathtt{b}) &= \mathtt{b}^5(\mathtt{ab})^{k-6}\mathtt{a},\\
\beta^\star(\mathtt{a}^2\mathtt{b}) &= \mathtt{aaba}^{2k-8},\\
\beta^\star(\mathtt{ab}) &= \mathtt{b}^{k-2}\mathtt{\$aba}^{2k-6},\\
\beta^\star(\mathtt{ba}) &= \mathtt{ba}^{k-5}\mathtt{bbbab}^{k-5}\mathtt{ab}^{k-2}\mathtt{a},\\
\beta^\star(\mathtt{b}^j\mathtt{a}) &= \mathtt{bab}^{2k-2j-2}\mathtt{a} \textit{ for all } 2 \leq j \leq k-1 \textit{ and}\\
\beta^\star(\mathtt{b}^k\mathtt{a}) &= \mathtt{a}.
\end{aligned}$$

Hence, $\mathrm{BWT}(w_k\mathtt{a\$}) = \beta^\star(\mathtt{\$})\cdot(\prod_{i=2}^{k-1}\beta^\star(\mathtt{a}^{k-i}\mathtt{b}))\cdot\beta^\star(\mathtt{b\$})\cdot(\prod_{i=1}^{k}\beta^\star(\mathtt{b}^i\mathtt{a}))$. *Moreover, it holds that* $r(w_k\mathtt{a\$}) = 8k-16$.

Proof We obtain $\mathrm{BWT}(w_k\mathtt{a\$}) = \mathtt{a}\mathrm{BWT}(w_k\mathtt{\$})$ by applying Proposition 36 to the words $w_k\mathtt{a\$}$ and $w_k\mathtt{\$}$, and we already know the structure of $\mathrm{BWT}(w_k\mathtt{\$})$ by Lemma 38. □

Proposition 42 *There exists an infinite family of words such that: (i)* $r_{\$}(w\mathtt{b}) - r_{\$}(w) = \Theta(\sqrt{n})$*; (ii)* $r_{\$}(\widehat{w}) - r_{\$}(w) = \Theta(\sqrt{n})$*; (iii)* $r_{\$}(\widehat{w}\mathtt{a}) - r_{\$}(w) = \Theta(\sqrt{n})$.

Proof Such a family is composed of the words $w_k\mathtt{b}$ with $k > 5$. The proof follows from Lemma 38, Lemma 39, Lemma 40, Lemma 41, and Observation 2. □

5.4 The Relationship Between r and $r_\$$

Now we address the differences between the measures r and $r_\$$. In fact, not only are the measures r and $r_\$$ not equal over the same input, but they may differ by a $\Theta(\log n)$ multiplicative factor, or by a $\Theta(\sqrt{n})$ additive factor.

Proposition 43 *There exists an infinite family of words v such that $r_\$(v)/r(v) = \Theta(\log n)$, where $n = |v|$.*

Proof The family consists of the reverse of the Fibonacci words of odd order. Let $v = \mathrm{rev}(s)$, with s a Fibonacci word of odd order $2k+1$. Since s is a standard word, $r(s) = 2$. Moreover, its reverse v is a conjugate and thus $\mathrm{BWT}(v) = \mathrm{BWT}(s)$, implying that also $r(v) = 2$. Let $v' = v\$$. Since $\$ < \mathtt{a}$, by Proposition 2 it follows that $r(v') \in \{2k+2, 2k+3\}$. Altogether, $r_\$(v)/r(v) \leq \frac{2k+3}{2} = \Theta(k) = \Theta(\log n)$.□

Proposition 44 *There exists an infinite family of words w such that $r_\$(w) - r(w) = \Theta(\sqrt{n})$, where $n = |w|$.*

Proof The family consists of the words w_k for all $k > 5$, defined in Section 4. From Proposition 30 and Lemma 38, it holds $r_\$(w_k) - r(w_k) = 2k - 4$. By Observation 2, it holds $r_\$(w_k) - r(w_k) = \Theta(\sqrt{n})$. □

6 Conclusion

In this paper, we studied how a single edit operation on a word (insertion, deletion or substitution of a character) can affect the number of runs r of the BWT of the word. Our contribution is threefold. First, we prove that $\Omega(\log n)$ is a lower bound for all three edit operations, by exhibiting infinite families of words for which each edit operation can increase the number of runs by a multiplicative $\Theta(\log n)$ factor. Since for all of these families, $r = \mathcal{O}(1)$, this also proves that the upper bound $\mathcal{O}(\log n \log r)$ given in [1] is tight in the case of $r = \mathcal{O}(1)$, for each of the three edit operations. Secondly, we improved the best known lower bound of $\Omega(\log n)$ for the additive sensitivity of r [1, 19], by giving an infinite family of words on which insertion, deletion, and substitution of a character can increase r by a $\Theta(\sqrt{n})$ additive factor. Finally, we put in relation the two common variants of the number of runs of the BWT, which we denote as r resp. $r_\$$. The latter, $r_\$$, is the variant used in articles on string data structures and compression, which assumes that each word is terminated by an end-of-string symbol; for the variant r commonly used in the literature on combinatorics on words, no such assumption is made.

Our work opens several roads of investigation. First, we ask whether there exist families of words with $r = \omega(1)$ for which edit operations can cause a multiplicative increase of $\Omega(\log n)$. In other words, is the bit catastrophe effect restricted to words on which the compression power of r is maximal?

Another interesting question is whether the upper bound $\mathcal{O}(r \log r \log n)$ from [1] for the additive sensitivity of r is tight. A weaker question, an answer to which would make a step in this direction, is whether there exists an infinite family with $r = \omega(1)$

on which one edit operation can cause an additive increase of $\omega(r)$ in the number of runs of the BWT.

Acknowledgements We thank two anonymous reviewers for their careful reading of our manuscript and several suggestions which helped improve the paper.

Author Contributions All authors contributed equally to the research and redaction of this project.

Funding Open access funding provided by Università degli Studi di Palermo within the CRUI-CARE Agreement. CU is funded by scholarship ANID-Subdirección de Capital Humano/Doctorado Nacional/2021-21210580, ANID, Chile.
ZsL, GR, and MS are partially funded by the MUR PRIN Project "PINC, Pangenome INformatiCs: from Theory to Applications" (Grant No. 2022YRB97K), and by the INdAM - GNCS Project CUP_E53C23001670001. MS is partially supported by the project "ACoMPA - Algorithmic and Combinatorial Methods for Pangenome Analysis" (CUP B73C24001050001) funded by the NextGeneration EU programme PNRR ECS00000017 Tuscany Health Ecosystem (Spoke 6).
SI is partially funded by JSPS KAKENHI grant numbers JP20H05964, JP23K24808, and JP23K18466.

Declarations

Conflicts of Interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akagi, T., Funakoshi, M., Inenaga, S.: Sensitivity of string compressors and repetitiveness measures. Inf. Comput. **291**, 104999 (2023)
2. Badkobeh, G., Bannai, H., Köppl, D.: Bijective BWT based compression schemes. In: Proceedings of 31st International Symposium on String Processing and Information Retrieval (SPIRE 2024), Volume 14899 of Lecture Notes in Computer Science, Cham, Springer, pp. 16–25 (2024)
3. Bannai, H., Gagie, T., T. I.: Refining the r-index. Theor. Comput. Sci. **812**, 96–108 (2020)
4. Berstel, J., de Luca, A.: Sturmian words, Lyndon words and trees. Theor. Comput. Sci. **178**(1–2), 171–203 (1997)
5. Biagi, E., Cenzato, D., Zs. Lipták, and G. Romana,: On the number of equal-letter runs of the bijective burrows-wheeler transform. In: Castiglione, G., Sciortino, M. (Eds.), Proceedings of the 24th Italian Conference on Theoretical Computer Science, Palermo, Italy, September 13-15, 2023, Volume 3587 of CEUR Workshop Proceedings, Aachen, pp 129–142. CEUR-WS.org (2023)
6. Boucher, C., Cenzato, D., Lipták, Z., Rossi, M., Sciortino, M.: Computing the original ebwt faster, simpler, and with less memory. In: Proceedings of 28th International Symposium on String Processing and Information Retrieval (SPIRE 2021), Volume 12944 of Lecture Notes in Computer Science, Cham, Springer, pp. 129–142 (2021)
7. Boucher, C., Cenzato, D., Zs. Lipták, M. Rossi, and M. Sciortino.: *r*-indexing the eBWT. Inf. Comput. **298**, 105155 (2024)

8. Brlek, S., Frosini, A., Mancini, I., Pergola, E., Rinaldi, S.: Burrows-Wheeler Transform of words defined by morphisms. In: IWOCA, Volume 11638 of Lecture Notes in Computer Science, Cham, Springer, pp. 393–404 (2019)
9. Burrows, M., Wheeler, D.J.: A block-sorting lossless data compression algorithm. Technical report, DIGITAL System Research Center (1994)
10. Castiglione, G., Restivo, A., Sciortino, M.: Circular sturmian words and Hopcroft's algorithm. Theor. Comput. Sci. **410**(43), 4372–4381 (2009)
11. Castiglione, G., Restivo, A., Sciortino, M.: On extremal cases of Hopcroft's algorithm. Theor. Comput. Sci. **411**(38–39), 3414–3422 (2010)
12. Cenzato, D., Lipták, Z.: A survey of BWT variants for string collections. Bioinformatics **40**(7), btae333 (2024)
13. de Luca, A.: Sturmian words: structure, combinatorics, and their arithmetics. Theor. Comput. Sci. **183**(1), 45–82 (1997)
14. de Luca, A., Mignosi, F.: Some combinatorial properties of Sturmian words. Theor. Comput. Sci. **136**(2), 361–285 (1994)
15. Ferragina, P., Manzini, G.: Opportunistic data structures with applications. In: 41st Annual Symposium on Foundations of Computer Science, FOCS 2000, 12-14 November 2000, Redondo Beach, California, USA, Los Alamitos, CA, IEEE Computer Society, pp. 390–398 (2000)
16. Frosini, A., Mancini, I., Rinaldi, S., Romana, G., Sciortino, M.: Logarithmic equal-letter runs for BWT of purely morphic words. In: Developments in Language Theory - 26th International Conference, DLT 2022, Tampa, FL, USA, May 9-13, 2022, Proceedings, Volume 13257 of Lecture Notes in Computer Science, Cham, Springer, pp. 139–151 (2022)
17. Gagie, T., Navarro, G., Prezza, N.: Optimal-time text indexing in BWT-runs bounded space. In: Czumaj, A. (Ed.), Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018, Philadelphia, PA, SIAM, pp. 1459–1477 (2018)
18. Gagie, T., Navarro, G., Prezza, N.: Fully functional suffix trees and optimal text searching in BWT-runs bounded space. J. ACM **67**(1), 2:1-2:54 (2020)
19. Giuliani, S., Inenaga, S., Zs. Lipták, N. Prezza, M. Sciortino, and A. Toffanello,: Novel results on the number of runs of the Burrows-Wheeler-Transform. In: Bures, T., Dondi, R., Gamper, J., Guerrini, G., Jurdzinski, T., Pahl, C., Sikora, F., Wong, P.W.H. (eds.) SOFSEM 2021: Theory and Practice of Computer Science - 47th International Conference on Current Trends in Theory and Practice of Computer Science, SOFSEM 2021, Bolzano-Bozen, Italy, January 25–29, 2021, Proceedings. Lecture Notes in Computer Science, Cham, vol. 12607, pp. 249–262. Springer (2021)
20. Giuliani, S., Inenaga, S., Zs. Lipták, G. Romana, M. Sciortino, and C. Urbina,: Bit catastrophes for the Burrows-Wheeler Transform. In: Drewes, F., Volkov, M. (eds.) Developments in Language Theory - 27th International Conference, DLT 2023, Umeå, Sweden, June 12–16, 2023, Proceedings. Lecture Notes in Computer Science, Cham, vol. 13911, pp. 86–99. Springer (2023)
21. Kempa, D., Kociumaka, T.: Resolution of the Burrows-Wheeler transform conjecture. Commun. ACM **65**(6), 91–98 (2022)
22. Knuth, D.E., Morris, J.H., Jr., Pratt, V.R.: Fast pattern matching in strings. SIAM J. Comput. **6**(2), 323–350 (1977)
23. Kociumaka, T., Navarro, G., Prezza, N.: Toward a definitive compressibility measure for repetitive sequences. IEEE Trans. Inf. Theory **69**(4), 2074–2092 (2023)
24. Lagarde, G., Perifel, S.: Lempel-Ziv: a "one-bit catastrophe" but not a tragedy. In: Czumaj, A. (Ed.), Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018, Philadelphia, PA, SIAM, pp. 1478–1495 (2018)
25. Lam, T.W., Li, R., Tam, A., Wong, S.C.K., Wu, E., Yiu, S.: High throughput short read alignment via bi-directional BWT. In: 2009 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2009, Washington, DC, USA, November 1-4, 2009, Proceedings, Los Alamitos, CA, IEEE Computer Society, pp. 31–36 (2009)
26. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**(4), 357–359 (2012)
27. Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**(3), R25 (2009)
28. Li, H., Durbin, R.: Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinform. **26**(5), 589–595 (2010)

29. Lothaire, M.: Algebraic Combinatorics on Words. Cambridge University Press, Cambridge (2002)
30. Mäkinen, V., Navarro, G.: Succinct Suffix Arrays based on Run-Length Encoding. In: Apostolico, A., Crochemore, M., Park, K. (Eds.), Combinatorial Pattern Matching, 16th Annual Symposium, CPM 2005, June 19-22, 2005, Jeju Island, Korea, Proceedings, Volume 3537 of Lecture Notes in Computer Science, Cham, Springer, pp. 45–56 (2005)
31. Mäkinen, V., Navarro, G.: Succinct suffix arrays based on run-length encoding. Nord. J. Comput. **12**(1), 40–66 (2005)
32. Mantaci, S., Restivo, A., Rosone, G., Sciortino, M.: An extension of the Burrows-Wheeler Transform. Theor. Comput. Sci. **387**(3), 298–312 (2007)
33. Mantaci, S., Restivo, A., Rosone, G., Sciortino, M., Versari, L.: Measuring the clustering effect of BWT via RLE. Theor. Comput. Sci. **698**, 79–87 (2017)
34. Mantaci, S., Restivo, A., Sciortino, M.: Burrows-Wheeler transform and Sturmian words. Inf. Process. Lett. **86**(5), 241–246 (2003)
35. Navarro, G.: Indexing highly repetitive string collections, part I: repetitiveness measures. ACM Comput. Surv. **54**(2), 29:1-29:31 (2022)
36. Rosone, G., Sciortino, M.: The Burrows-Wheeler Transform between Data Compression and Combinatorics on Words. In: Bonizzoni, P., Brattka, V., Löwe, B. (Eds.), The Nature of Computation. Logic, Algorithms, Applications - 9th Conference on Computability in Europe, CiE 2013, Milan, Italy, July 1-5, 2013. Proceedings, Volume 7921 of Lecture Notes in Computer Science, Cham, Springer, pp. 353–364 (2013)
37. Sciortino, M., Zamboni, L.Q.: Suffix Automata and Standard Sturmian Words. In: Harju, T., Karhumäki, J., Lepistö, A. (eds.) Developments in Language Theory, 11th International Conference, DLT 2007, Turku, Finland, July 3–6, 2007, Proceedings. Lecture Notes in Computer Science, Cham, vol. 4588, pp. 382–398. Springer (2007)
38. Seward, J.: https://sourceware.org/bzip2/manual/manual.html (1996)
39. Vasimuddin, M., Misra, S., Li, H., Aluru, S.: Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2019, Rio de Janeiro, Brazil, May 20-24, 2019, Los Alamitos, CA, IEEE Computer Society, pp. 314–324 (2019)
40. Ziv, J., Lempel, A.: A universal algorithm for sequential data compression. IEEE Trans. Inf. Theory **23**(3), 337–343 (1977)
41. Ziv, J., Lempel, A.: Compression of individual sequences via variable-rate coding. IEEE Trans. Inf. Theory **24**(5), 530–536 (1978)