Arithmetics on Suffix Arrays of Fibonacci Words

Dominik Köppl and Tomohiro I

Department of Computer Science, TU Dortmund, Germany
dominik.koeppl@tu-dortmund.de, tomohiro.i@cs.tu-dortmund.de

Abstract. We study the sequence of Fibonacci words and some of its derivatives with respect to their suffix array, inverse suffix array and Burrows-Wheeler transform based on the respective suffix array. We show that the suffix array is a rotation of its inverse under certain conditions, and that the factors of the LZ77 factorization of any Fibonacci word yield again similar characteristics.

1 Introduction

The sequence of Fibonacci words is one of the best studied set of strings in the field of combinatorics. A Fibonacci word, composed by the concatenation of its predecessor with its pre-predecessor, excels at many interesting properties regarding factorizations [18], powers [16], fractals [15], entropy [7] and palindromes [4]. The sequence is often used as a testbed for algorithms, since they are a representative for some worst case scenarios [8]. Regarding benchmarks, it is beneficial to know the shape of the considered data structures when they are applied to a Fibonacci word; studying the combinatorial properties may help understanding experimental results, or may even lead to designing new algorithms.

We study properties of Fibonacci words and of some derivatives with respect to its suffix array (SA) [13]. The SA induces the structure of its inverse and of the SA-based Burrows-Wheeler transform (BWT) [2, 10]. Under certain conditions, the SA is a rotation, a reversed rotation, or the copy of its inverse.

With this insight, it is easy to reason about complex data structures like compressed suffix trees [19], the FM-Index [5] or LZ77-based self-indexes [6].

2 Related Work

Regarding suffix data structures, Rytter [17] considered building a directed acyclic word graph and a suffix tree on the j-th Fibonacci word. By some properties of the Fibonacci words, Rytter showed an easy way to modify both data structures to match with the $(j+1)$-th word. Considering BWT based on rotations, Mantaci et al. [14] discovered that the BWT rearranges any Fibonacci word in a block of consecutive b's, followed by a block of consecutive a's. More generally, Mantaci et al. [14] and Simpson and Puglisi [20] gave a general theorem regarding this special shape of the BWT applied to a class of binary strings, so called standard words, to which the Fibonacci words belong. Christodoulakis et al. [3] proposed a constant time algorithm for querying different properties on the BWT's rotations. Without any delimiting, unique character at the string's end, the BWT defined on rotations and the BWT based on the SA (may) differ. In fact, the BWT based on the SA of F_n (for any odd n) does not transform the string into two homogenous blocks. Another result with respect to rotations is done by Droubay [4]: they showed for every F_n with $n \bmod 3 \neq 0$ that there exists exactly one k such that the k-th rotation of F_n is a palindrome.

Since some current, popular indexing strategies perform compression on the text (e.g., [6]), we further consider the LZ77 factorization [23]. In the special case of the Fibonacci words, Berstel and Savelli [1] pointed out that the LZ77 factorization coincides with the palindromic factorization studied by Wen and Wen [22]. We will show that our results apply analogously to the LZ77 factors.

3 Preliminaries

Let Σ denote an ordered alphabet. An element in Σ^* is called a ***string***. For any string T, let $|T|$ denote the length of T. The string of length zero is denoted by ϵ. Σ^* forms with ϵ and the concatenation $\Sigma^* \times \Sigma^* \to \Sigma^*, (u, v) \mapsto uv$ a free monoid. For any $1 \leq i \leq |T|$, $T[i]$ denotes the i-th character of T. When $T \in \Sigma^*$ is represented by the concatenation of $x, y, z \in \Sigma^*$, i.e., $T = xyz$, then x, y and z are called a ***prefix***, ***substring*** and ***suffix*** of T, respectively. For any $1 \leq i \leq j \leq |T|$, a substring of T starting at i and ending at j is denoted by $T[i..j]$. Especially, a suffix starting at position i of T is denoted by $T[i..]$. For any $x, y \in \Sigma^*$, let $\mathsf{lcp}(x, y)$ denote the length of the longest common prefix of x and y.

The ***lexicographical order*** is denoted by $<\subset \Sigma^* \times \Sigma^*$, i.e., $x < y$ iff (either) x is a proper prefix of y, or $l := \mathsf{lcp}(x, y)$ is less than $\min(|x|, |y|)$ and $x[l + 1] < y[l + 1]$. We use another ordering $\lessdot$ for that $x \lessdot y$ iff the latter condition holds. The ordering $\lessdot$ is finer than $<$. If $x \lessdot y$, then $xu \lessdot yv$ holds for any $u, v \in \Sigma^*$. For instance, $a < aa$ and $a \not\lessdot aa$ since a is a prefix of aa. Appending b to both strings flips the lexicographic order to $ab > aab$. Taking $aa \lessdot ab$ as an example, appending characters to both strings does not affect their ordering.

The ***inverse*** R^{-1} of an array R is an array with the same length for that $R^{-1}[R[i]] = i$ holds for every $1 \leq i \leq |R|$; the inverse of R exists iff R is a ***permutation*** over $\{1, \ldots, |R|\}$, i.e., $\lambda.j \mapsto R[j]$ is an injective endomorphism. The ***suffix array*** SA_T of a string T is an array of length $|T|$ such that $T[\mathsf{SA}_T[i]..] < T[\mathsf{SA}_T[i + 1]..]$ for every $1 \leq i < |T|$. Since SA_T is a permutation, its inverse (i.e., the ***inverse suffix array***) exists, and is denoted by ISA_T.

Like in common literature, we let arrays start at position one (not zero); therefore we will modify the modulo operator not to map any value to zero. For this purpose, we define the modulo operator on the natural numbers by $\bmod n : \mathbb{N} \to \{1, \ldots, n\} \subset \mathbb{N}$, $m \bmod n := \ m - n \bmod n$ if $m > n$, $m \bmod n := m$ otherwise, for $n, m \in \mathbb{N}$. Naturally, our results can also be applied to the standard modulo operator when taking arrays of the form $[0..n - 1]$, instead of $[1..n]$.

We call the array ψ_T with $\psi_T[i] := \mathsf{ISA}_T[\mathsf{SA}_T[i] - 1 \mod |T|]$ for every $1 \leq i \leq |T|$ the ***last-to-front*** mapping of T.

A permutation S is called a ***rotation*** of R iff there exists exactly one $k \in \{1, \ldots, |R|\}$ such that $R[i] = S[(k+i) \mod |R|]$. A permutation S is called a ***reversed rotation*** of R iff there exists one $k \in \{1, \ldots, |R|\}$ such that $R[i] = S[(k - i) \mod |R|]$. In both cases, we call k the ***shift*** of S. If S is a rotation of R and there exists $1 \leq i \leq n$ such that $R[i] = S[i]$, then $S = R$.

Let $\Sigma_2 = \{a, b\}$ be a binary alphabet with $a < b$. The ***complementation*** $\bar{\cdot} : \Sigma_2^* \to \Sigma_2^*$ complements a string, i.e., $\bar{T}[i] = \ a$ if $T[i] = b$, $\bar{T}[i] = b$ if $T[i] = a$.

Definition 1. *The n-th* **Fibonacci word** $F_n \in \Sigma_2^*$ *($n \in \mathbb{N}$) is defined by $F_n = \ b$ if $n = 1$, $F_n = a$ if $n = 2$, $F_n = F_{n-1}F_{n-2}$ otherwise. The sequence of lengths $f_n := |F_n|$ form the* **Fibonacci numbers**. *The sequence $\{\bar{F_n}\}_{n\in\mathbb{N}}$ is called* **rabbit sequence** *[7] and sometimes confused with the Fibonacci words*

Sequence	$n \geq 4$	SA $\leftrightarrow$ ISA	shift	BWT
F_n (Def. 1)	even	rotation	$f_{n-2}+1$	$b^{f_{n-2}}a^{f_{n-1}}$
Z_n (Def. 3)	even	rotation	$f_{n-2}+1$	$b^{f_{n-2}}a^{f_{n-1}-1}b$
$B_n := \beta F_n$	even	equal	0	$b^{f_{n-2}}\beta a^{f_{n-1}}$
$D_n := \bar{F}_n c$	even	equal	0	$b^{f_{n-1}-1}ca^{f_{n-2}}b$
$C_n := F_n c$	odd	reversed rotation	f_n	$b^{f_{n-2}-1}ca^{f_{n-1}}b$

Table 1. For each string T of a given sequence, we show the relationship between SA_T and ISA_T, as well as the shape of BWT_T. β is a variable character with $\beta \geq b$, and c is a character with $c > b$.

(e.g., see [11]). The ending of the n-th Fibonacci word ($n \geq 3$) is given by $\delta_n := F_n[f_n - 1..f_n]$ such that $\delta_n =$ ba if n is even, $\delta_n = ab$ if n is odd.

A ***factorization*** partitions T into z substrings $T = w_1 \cdots w_z$. These substrings are called ***factors***. In particular, we have:

Definition 2 ([23]). *A factorization $w_1 \cdots w_z = T$ is called the* **LZ77 factorization** *of T iff w_x is the shortest prefix of $w_x \cdots w_z$ that occurs exactly once in $w_1 \cdots w_x$.*

Definition 3 ([22, 12, 1]). *The n-th* **singular word** *is defined as $Z_n := \overline{F_n[f_n]}F_n[1..f_n - 1]$. Alternatively, Z_n can be written as $Z_n =$ a if $n = 1$, $Z_n = b$ if $n = 2$, $Z_n = aa$ if $n = 3$, $Z_n = Z_{n-2}Z_{n-3}Z_{n-2}$ otherwise.*

Lemma 1. *For any $n \geq 3$, $F_n = Z_1 \cdots Z_{n-2}\gamma$ is the LZ77 factorization of F_n, where $\gamma := \delta_n[1]$, i.e., $\gamma =$ b if n is odd, $\gamma = a$ if n is even.*

Proof.
- **[basis]** $F_3 = Z_1 b = ab$, $F_4 = Z_1 Z_2 a = aba$.
- **[hypothesis]** Assume the claim holds for $n-1$ and $n-2$.
- **[induction proof]** Let $\gamma := F_{n-2}[f_{n-2}] = F_n[f_n]$. Then $Z_{n-2} = \bar{\gamma}F_{n-2}[1..f_{n-2}-1] = \bar{\gamma}Z_1 \cdots Z_{n-4}$, and $F_n = Z_1 \cdots Z_{n-3}\bar{\gamma}Z_1 \cdots Z_{n-4}\gamma = Z_1 \cdots Z_{n-2}\gamma$.

□

Table 1 gives a summary of the properties shown in this paper.

Remark 1. The arithmetic progression that characterizes SA and ISA is not restricted to the family of Fibonacci-like strings. For example, the sequences $S_1 = abaa$, $S_n = aS_{n-1}a$ and $E_0 = bb, E_1 = bbab, E_n = b^{n+1}a^n b$ have a suffix array that is a reverse rotation of its inverse. Instances of both sequences are depicted in Table 3.

4 The Suffix Array and its Inverse

We examine the suffix array structure of each sequence considered in Table 1. Besides this, we are interested in revealing some relationship between SA and ISA. Some examples are depicted in Table 2. Lemma 3 gives us some rules that determine whether a specific array is a rotation or reversed rotation of its inverse. For our sequences, Lemma 5 shows that the suffix array has the form of the array dealt in Definition 4 for some certain n.

i	1	2	3	4	5	6	7	8
$F_6[i]$	a	b	a	a	b	a	b	a
$\mathsf{SA}_{F_6}[i]$	8	3	6	1	4	7	2	5
$\mathsf{ISA}_{F_6}[i]$	4	7	2	5	8	3	6	1
Rotation Shift: 4								

i	1	2	3	4	5	6	7	8
$Z_6[i]$	b	a	b	a	a	b	a	b
$\mathsf{SA}_{Z_6}[i]$	4	7	2	5	8	3	6	1
$\mathsf{ISA}_{Z_6}[i]$	8	3	6	1	4	7	2	5
Rotation Shift: 4								

i	1	2	3	4	5	6	7	8	9
$(\bar{F}_6c)[i]$	b	a	b	b	a	b	a	b	c
$\mathsf{SA}_{\bar{F}_6c}[i]$	5	2	7	4	1	6	3	8	9
$\mathsf{ISA}_{\bar{F}_6c}[i]$	5	2	7	4	1	6	3	8	9

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$(F_7c)[i]$	a	b	a	a	b	a	b	a	a	b	a	a	b	c
$\mathsf{SA}_{F_7c}[i]$	8	3	11	6	1	9	4	12	7	2	10	5	13	14
$\mathsf{ISA}_{F_7c}[i]$	5	10	2	7	12	4	9	1	6	11	3	8	13	14
Reverse Rotation Shift: 13														

i	1	2	3	4	5	6	7	8	9
$(\beta F_6)[i]$	β	a	b	a	a	b	a	b	a
$\mathsf{SA}_{\beta F_6}[i]$	9	4	7	2	5	8	3	6	1
$\mathsf{ISA}_{\beta F_6}[i]$	9	4	7	2	5	8	3	6	1

i	1	2	3	4	5	6	7	8	9	10	11	12	13
F_7	a	b	a	a	b	a	b	a	a	b	a	a	b
SA_{F_7}	11	8	3	12	9	6	1	4	13	10	7	2	5
ISA_{F_7}	7	12	3	8	13	6	11	2	5	10	1	4	9
BWT_{F_7}	b	b	b	a	a	b	b	a	a	a	a	a	a

i	1	2	3	4	5	6	7	8	9
F_6c	a	b	a	a	b	a	b	a	c
SA_{F_6c}	3	1	4	6	8	2	5	7	9
ISA_{F_6c}	2	6	1	3	7	4	8	5	9
BWT_{F_6c}	b	c	a	b	b	a	a	a	a

Table 2. Instances of the string sequences considered in Table 1. We additionally examine F_7 that does not show any of the attractive properties we study. Neither F_7 nor F_6c possesses any interesting properties we focus on.

i	1	2	3	4	5	6	7	8
$S_2[i]$	a	a	a	b	a	a	a	a
$\mathsf{SA}_{S_2}[i]$	8	7	6	5	1	2	3	4
$\mathsf{ISA}_{S_2}[i]$	5	6	7	8	4	3	2	1
$\mathsf{BWT}_{S_2}[i]$	a	a	a	b	a	a	a	a
Rev. Rot. Shift: 5								

i	1	2	3	4	5	6	7	8
$E_3[i]$	b	b	b	b	a	a	a	b
$\mathsf{SA}_{E_3}[i]$	5	6	7	8	4	3	2	1
$\mathsf{ISA}_{E_3}[i]$	8	7	6	5	1	2	3	4
$\mathsf{BWT}_{E_3}[i]$	b	a	a	a	b	b	b	b
Rev. Rot. Shift: 5								

Table 3. Instances of the string sequences given in Remark 1. Both instances have an SA that is reverse rotated to its inverse.

Definition 4. *Let R be an array of integers with length $n \in \mathbb{N}$. We call R* **arithmetic progressed** *iff there exists $m < n$ and $q \in \{1, \ldots, n\}$ such that $R[i] = \; q \;$ if $i = 1$, $R[i] = (R[i-1] + m) \mod n \;$ if $i > 1$, for $1 \leq i \leq n$.*

Lemmas 2 to 4 consider an array R that is arithmetic progressed. Let n, m and q be defined as in Definition 4.

Lemma 2. *R is a permutation iff $\gcd(m, n) = 1$.*

Proof. By construction, R is an endomorphism. Let us take any $r \in \{1 \ldots, n-1\}$. By $R[(i + r) \mod n] = (R[i] + rm) \mod n$ we see that

$$\begin{aligned} \gcd(n, m) = 1 &\Leftrightarrow rm \mod n \neq n \; \forall 1 \leq r \leq n-1 \\ &\Leftrightarrow (R[i] + rm) \mod n \neq R[i] \; \forall 1 \leq i \leq n \text{ and } \forall 1 \leq r \leq n-1. \end{aligned}$$

□

Lemma 3. *Considering the inverse R^{-1} of R, the following properties hold:*

a) The array R^{-1} *is a rotation of* R *with shift* $(q+(q-1)m-1) \mod n$ *if and only if* $m^2 \mod n = 1$ *and* $\gcd(m,n)=1$ *holds.*

b) If R^{-1} *is a rotation of* R *and* $q \in \{1, m\}$*, then* $R^{-1} = R$.

c) The array R^{-1} *is a reversed rotation of* R *with shift* $(q+(q-1)m+1) \mod n$ *if and only if* $m^2 \mod n = n-1$ *and* $\gcd(m,n)=1$ *holds.*

Proof. a) Let $x := R[i]$ for an arbitrary, fixed $1 \leq i \leq n$. Then $R[(i+m) \mod n] = (x+m^2) \mod n$. We conclude that $R^{-1}[x] = i$ and $R^{-1}[(x+m^2) \mod n] = (i+m) \mod n$ holds. Now we yield the equivalence

$$R^{-1}[(x+1) \mod n] = (i+m) \mod n \Leftrightarrow m^2 \mod n = 1.$$

Since $R[1] = q$, the shift is $R[q] - R^{-1}[q] \mod n = (q+(q-1)m-1) \mod n$.

b) If $q=1$, then $R[1]=1$; hence 1 is a fix point. If $q=m$, then $R[m+1] = (q+m^2) \mod n = m+1$; hence $m+1$ is a fix point.

c) Let x, i be defined as in proof of Item a). Then we yield the equivalence

$$R^{-1}[(x-1) \mod n] = (i+m) \mod n \Leftrightarrow m^2 \mod n = n-1.$$

Since $R[1] = q$, the shift is $R[q] + R^{-1}[q] \mod n = (q+(q-1)m+1) \mod n$.

□

Lemma 4. *The last-to-front mapping* $\psi_R[i] := R^{-1}[R[i]-1 \mod n]$ *shows the following characterizations:*

a) If R^{-1} *is a rotation of* R*, then* $\psi_R[i] = (i-m) \mod n$.

b) If R^{-1} *is a reversed rotation of* R*, then* $\psi_R[i] = (i+m) \mod n$.

Proof. We follow the observations in Lemma 3.

a) Let k denote the shift of R^{-1}. Then $\psi_R[i] = R^{-1}[R[i]-1 \mod n] = R^{-1}[q+(i-1)m-1 \mod n] = R[q+(i-1)m-1-k \mod n] = q+(q+(i-1)m-k-2)m \mod n = i-m \mod n$.

b) Let k denote the shift of R^{-1}. Then $\psi_R[i] = R^{-1}[R[i]-1 \mod n] = R^{-1}[q+(i-1)m-1 \mod n] = R[k-q-(i-1)m+1 \mod n] = q+km-qm-(i-1)m^2 \mod n = q+qm+(q-1)m^2+m-qm-(i-1)m^2 \mod n = i+m \mod n$.

□

The following, well-known properties of the Fibonacci numbers allow us to apply Lemmas 2 to 4 to the suffix array and the last-to-front mapping of some instances of the string sequences under consideration:

Lemma 5 ([21, 9]). *The following statements hold:*

- $\gcd(f_n, f_{n-1}) = 1$ *for* $n \in \mathbb{N}$.
- *For* $n > 1$ *even,* $f_{n-1}^2 \mod f_n = 1$ *holds.*
- *For* $n > 1$ *odd,* $f_{n-1}^2 \mod f_n = n-1$ *holds.*
- *Since* $(n-m)^2 \mod n = m^2 \mod n$ *for any* $m, n \in \mathbb{N}$*, we can exchange* f_{n-1} *by* f_{n-2} *in the items above.*

Since $\mathsf{SA}_{F_1}, \mathsf{SA}_{F_2}$ and SA_{F_3} are the identity, we focus on the strings F_n with $n \geq 4$.

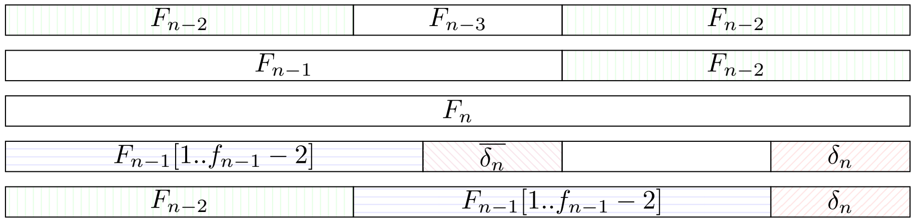

Fig. 1. Overview over the different split-ups considered for F_n with $n \geq 4$. The lower part is shown in proof of Lemma 7 and used by Lemma 10.

Lemma 6 (Christodoulakis et al. [3, Lemma 2.8]). *For* $n > 3$, $F_n = F_{n-2}F_{n-3} \cdots F_2\delta_n$.

Lemma 7. *For* $n \geq 4$ *and* $1 \leq i < f_{n-1}$, *we have*

$F_n[i..] \lessdot F_n[i + f_{n-2}..]$ *if* n *is even,* $F_n[i..] \gtrdot F_n[i + f_{n-2}..]$ *if* n *is odd.*

Proof. It follows from Lemma 6 that $F_n = F_{n-2}F_{n-3} \cdots F_2\delta_n$ and $F_n[1..f_{n-1}] = F_{n-1}$ and $F_{n-1} = F_{n-3}F_{n-4} \cdots F_2\overline{\delta_n}$. So $\mathsf{lcp}\,(F_n[1..], F_n[1 + f_{n-2}..]) = f_{n-1} - |\delta_n|$. For any $1 \leq i < f_{n-1}$, the order $\lessdot$ of $F_n[i..]$ and $F_n[i + f_{n-2}..]$ is determined by comparing $F_n[f_{n-1} - 1] = \overline{\delta_n}[1]$ with $F_n[f_n - 1] = \delta_n[1]$. □

Lemma 8. *For* $n \geq 4$ *even,* $F_n[f_n..] < F_n[f_n + f_{n-2} \mod f_n..] < F_n[f_n + 2f_{n-2} \mod f_n..] < \ldots < F_n[f_n + (f_n - 1)f_{n-2} \mod f_n..]$.

Proof. The conclusion of the inequations is divided up into two intervals and a starting position:

- It follows by Lemma 7 that $F_n[i..] \lessdot F_n[i + f_{n-2}..]$ for all $1 \leq i < f_{n-1}$.
- Since $F_n = F_{n-1}F_{n-2} = F_{n-2}F_{n-3}F_{n-2}$, $F_n[i..]$ is a prefix of $F_n[i - f_{n-1}..] = F_n[i + f_{n-2} \mod f_n..]$ for every $f_{n-1} < i \leq f_n$. Hence, $F_n[i..] < F_n[i + f_{n-2} \mod f_n..]$.
- By Lemma 6, $F_n[f_n..]$ is the lexicographically smallest suffix of F_n. Since f_n and f_{n-2} are coprime, there is a lexicographically increasing chain starting at $F_n[f_n..]$ that visits every suffix of F_n by a step of f_{n-2} (taking modulo f_n). The lexicographically largest suffix is $F_n[f_n + (f_n - 1)f_{n-2} \mod f_n..] = F_n[f_{n-1}..]$.

□

Theorem 1. *For* $n \in \mathbb{N}$ *even,* ISA_{F_n} *is a rotation of* SA_{F_n} *with a shift of* $f_{n-2} + 1$. SA_{F_n} *is given by* $\mathsf{SA}_{F_n}[i] = f_n$ *if* $i = 1$, $\mathsf{SA}_{F_n}[i] = (\mathsf{SA}_{F_n}[i-1] + f_{n-2}) \mod f_n$ *otherwise.*

Proof. The arithmetic characterization of SA_{F_n} follows directly from Lemma 8. By Lemma 3(a), ISA_{F_n} is a rotation of SA_{F_n}. □

Theorem 2. *Let* $B_n := \beta F_n$ *with a character* $\beta \geq b$. *For* $n \in \mathbb{N}$ *even,* ISA_{B_n} *is equal to* SA_{B_n}, *which is given by* $\mathsf{SA}_{B_n}[i] = f_n + 1$ *if* $i = 1$, $\mathsf{SA}_{B_n}[i] = (\mathsf{SA}_{B_n}[i-1] + f_{n-2})$ *otherwise.*

Proof. By Lemma 8 we know the lexicographical order of the suffixes $B_n[i..]$ for $i > 1$. It remains to pigeonhole $B_n[1..]$. Theorem 1 tells us that $F_n[f_{n-1}..]$ is the largest suffix of f_n. By transitivity it suffices to show that $F_n[f_{n-1}..] < B_n$: this is clear if $\beta > b$. Otherwise ($\beta = b$), $B_n[1..] = bF_n[1..]$ and $F_n[f_{n-1}..] = bF_n[1 + f_{n-1}..] = bF_{n-2}$. But we saw in proof of Lemma 8 that $F_{n-2} < F_n$, hence $F_n[f_{n-1}..] < B_n$. Together we get that the step is f_{n-2}.

We complete the arithmetic characterization of SA_{B_n} by showing a fix point:

$$\mathsf{SA}_{B_n}[f_{n-2}+2] = \left(f_{n-2}^2 + f_{n-2} + 1\right) \mod f_n = f_{n-2}+2,$$

where we used that $f_{n-2}^2 \mod f_n = 1$ by Lemma 5, and $|F_{m-2}| + 2 < |F_m|$ for every $m > 4$. Hence, by Lemma 3(b), $\mathsf{ISA}_{B_n}[2..f_n - 1]$ is equal to $\mathsf{SA}_{B_n}[2..f_n - 1]$. □

For the sequences C_n and D_n we have results similar to Lemma 8:

Corollary 1. *a) Let $C_n := F_n c$ with a character $c > b$. For $n \geq 5$ odd, $C_n[f_{n-1}..] < C_n[2f_{n-1} \mod f_n..] < \ldots < C_n[f_n f_{n-1} \mod f_n..] = C_n[f_n..]$.*

b) Let $D_m := \bar{F}_m c$ with a character $c > b$. For $m \geq 4$ even, $D_m[f_{m-1}..] < D_m[2f_{m-1} \mod f_m..] < \ldots < D_m[f_m f_{m-1} \mod f_m..] = D_m[f_m..]$.

Proof. We follow the steps of the proof of Lemma 8: It follows from Lemma 7 that $C_n[i..] \gtrdot C_n[i + f_{n-2}..]$ for all $1 \leq i < f_{n-1}$. Hence $C_n[i..] \lessdot C_n[i + f_{n-1} \mod f_n..]$ for all $f_{n-2} < i \leq f_n$. By the same argument, $\bar{F}_m[i..] \gtrdot \bar{F}_m[i + f_{m-2}..]$ for all $1 \leq i < f_{m-1}$, thus $D_m[i..] \lessdot D_m[i + f_{m-1} \mod f_m..]$ for all $f_{m-2} < i \leq f_m$. $F_n[i + f_{n-1}..]$ is a prefix of $F_n[i..]$ for every $1 \leq i \leq f_{n-2}$. So the order $C_n[i + f_{n-1}..] \gtrdot C_n[i..]$ is determined by comparing c with $F_{n-3}[1]$. Since $D_m = \bar{F}_{m-2}\bar{F}_{m-3}\bar{F}_{m-2}c$, $\bar{F}_m[i + f_{n-1}..]$ is a prefix of $\bar{F}_m[i..]$ for every $1 \leq i \leq f_{m-2}$. So the order $D_m[i + f_{n-1}..] \gtrdot D_m[i..]$ is determined by comparing c with $\bar{F}_{m-3}[1]$. So far, we have $C_n[i..] \lessdot C_n[i + f_{n-1} \mod f_n..]$ for every $1 \leq i < f_n$, and $D_m[i..] \lessdot D_m[i + f_{m-1} \mod f_m..]$ for every $1 \leq i < f_m$. Since f_n and f_{n-1} are coprime, there is a lexicographically increasing chain starting at $C_n[f_{n-1}..]$ that visits every suffix of C_n, except $C_n[f_n + 1..]$, by a step of f_{n-1} (taking modulo f_n); the same holds for D_m. □

Theorem 3. *Let $C_n := F_n c$ with a character $c > b$. For $n \in \mathbb{N}$ odd, $\mathsf{ISA}_{C_n}[1..f_n]$ is a reversed rotation of $\mathsf{SA}_{C_n}[1..f_n]$ with a shift of f_n. SA_{C_n} is given by $\mathsf{SA}_{C_n}[i] = f_n + 1$ if $i = f_n + 1$, $\mathsf{SA}_{C_n}[i] = f_n$ if $i = f_n$, $\mathsf{SA}_{C_n}[i] = (\mathsf{SA}_{C_n}[i+1] + f_{n-1}) \mod f_n$ if $1 \leq i < f_n$.*

Proof. Since $C_n[f_n + 1..] = c$ is the largest suffix of C_n, the arithmetic characterization of SA_{C_n} follows from Corollary 1(a). By Lemma 3(c), $\mathsf{ISA}_{C_n}[1..f_n]$ is a reversed rotation of $\mathsf{SA}_{C_n}[1..f_n]$. □

Theorem 4. *Let $D_n := \bar{F}_n c$ with a character $c > b$. For $n \in \mathbb{N}$ even, ISA_{D_n} is equal to SA_{D_n}; both are given by $\mathsf{SA}_{D_n}[i] = f_n + 1$ if $i = f_n + 1$, $\mathsf{SA}_{D_n}[i] = f_n$ if $i = f_n$, $\mathsf{SA}_{D_n}[i] = (\mathsf{SA}_{D_n}[i+1] - f_{n-2}) \mod f_n$ if $1 \leq i \leq f_n$.*

Proof. Since $D_n[f_n + 1..] = c$ is the largest suffix of D_n, the arithmetic characterization of SA_{D_n} follows from Corollary 1(b). By Lemma 3(b), ISA_{D_n} is equal to SA_{D_n}. □

Lemma 9. *For $n \geq 4$ even, $Z_n[1 + f_{n-2}..] < Z_n[1 + 2f_{n-2} \mod f_n..] < \ldots < Z_n[1 + f_n f_{n-2} \mod f_n..] = Z_n[1..]$, where Z_n is the n-th singular word, defined in Definition 3.*

Proof. Let $G := F_n[1..f_n - 1]$. For n even, $Z_n = bF_n[1..f_n - 1]$. Following the proof of Lemma 8 for G (Lemma 6 is still applicable for G since it depends on the $\delta_n[1]$-value, not on $\delta_n[2]$) yields $G[i..] \lessdot G[i + f_{n-2} \mod f_n..]$ for every $1 \leq i < f_{n-1} - 1$. Thus, $Z_n[i..] \lessdot Z_n[i + f_{n-2} \mod f_n..]$ for every $1 < i < f_{n-1}$. For the other i-values we consider that $Z_n = Z_{n-2}Z_{n-3}Z_{n-2}$; hence $Z_n[i + f_{n-1}..]$ is a prefix of $Z_n[i..]$ for

$T[\mathsf{SA}_T[i]]$	$a \ldots a$	$a \ldots a$	$b \ldots b$
$T[\mathsf{SA}_T[i]-1]$	$b \ldots b$	$a \ldots a$	$a \ldots a$
Blocks	ba-type	aa-type	ab-type

Table 4. We divide the suffixes of the considered strings in Section 5 into blocks of ba-,aa- and ba-type (to some extent). These types are arranged (mostly) like for the text T in this table.

every $1 \leq i \leq f_{n-2}$. So $Z_n[i..] < Z_n[i + f_{n-2} \mod f_n..]$ for every $f_{n-1} < i \leq f_n$. To sum up, there is a lexicographically increasing chain starting at $Z_n[1 + f_{n-1}..]$ that visits every suffix of Z_n by a step of f_{n-2} (taking modulo f_n). □

Theorem 5. *For $n \in \mathbb{N}$ even, ISA_{Z_n} is a rotation of SA_{Z_n} with a shift of $f_{n-2}+1$. SA_{Z_n} is given by* $\mathsf{SA}_{Z_n}[i] = \; f_{n-2}+1 \;$ *if* $i = 1$, $\mathsf{SA}_{Z_n}[i] = (\mathsf{SA}_{Z_n}[i-1] + f_{n-2}) \mod f_n$ *otherwise.*

Proof. The arithmetic characterization of SA_{Z_n} follows from Lemma 9. ISA_{Z_n} is a rotation of SA_{Z_n}, due to Lemma 3(a). □

5 Burrows-Wheeler Transform

In this section, we give a characterization of the BWT for the string sequences displayed by Table 1. We generate BWT_T of any string T by taking the preceding character of the suffix $T[\mathsf{SA}_T[i]..]$ while successively incrementing $1 \leq i \leq |T|$. Fortunately, the acquired results for SA_T can be directly applied for constructing BWT_T: Consider any $T \in \Sigma_2^*$ whose SA_T is a rotation (or reversed rotation) of its inverse. Further, consider that SA_T is arithmetic progressed; then the previous/next entry in SA_T is determined by a step of m or $n-m$ (we introduce m, n like in Definition 4). If the b-entries in T are distributed in such a way that we find a b at position $i+m$ or $i+n-m$ for each $1 \leq i \leq |T|$ with $T[i] = b$, then the suffixes $T[i..]$ that succeed a b ($= T[i-1]$) are aligned successively in SA_T, which means that the BWT_T generates a homogenous block of b's, like in Table 4. This stepping-characteristic is caught by the following

Lemma 10 (Rytter [17, Figure 2]). *For any $n \geq 4$, we have*

- $F_n[i] = F_n[i + f_{n-2}]$ *for any* $i \in \{1, \ldots, f_{n-1} - 2\}$, *and*
- $F_n[i] = F_n[i + f_{n-1}]$ *for any* $1 \leq i \leq f_{n-2}$.

Proof. By Lemma 6, $F_n = F_{n-2}F_{n-3} \cdots F_1\delta_n$. Also, $F_n[1..f_{n-1}] = F_{n-3}F_{n-4} \cdots F_1\overline{\delta_n}$. The second claim follows by splitting $F_n = F_{n-2}F_{n-3}F_{n-2}$. Figure 1 illustrates the proven properties. □

Theorem 6. *For n even, we get $\psi_{F_n}[i] = (i + f_{n-1}) \mod f_n$ and $\mathsf{BWT}_{F_n} = b^{f_{n-2}}a^{f_{n-1}}$.*

Proof. Because F_n does not contain the string bb, the only substrings of length two are aa, ab and ba. We will focus on the suffixes that start with an a that are preceded by a b. We call these of type ba; they are depicted in Table 4. If we show that these suffixes have successive numbers in the beginning of SA_{F_n}, we yield the claimed structure of BWT_{F_n}. We find these suffixes by tracking the chain given in the proof of Lemma 8. The chain starts at the smallest suffix $F_n[f_n..]$ and takes steps of length f_{n-2} (modulo f_n). Fortunately,

$F_n[f_n..]$ is exactly a ba-type suffix with $F_n[f_n - 1..] = \delta_n = ba$. By Lemma 10, the chain will visit iteratively the next ba-type suffix, until it accesses the ba-type suffix $\mathsf{SA}_{F_n}[f_{n-2}] = f_{n-1} + 1$. This is the last ba-type suffix: $\mathsf{SA}_{F_n}[f_{n-2} + 1] = 1$, and by definition of the BWT, the preceding character of the suffix $\mathsf{SA}_{F_n}[1..]$ is $F_n[f_n] = a$. Because F_n contains exactly f_{n-2} many b's, we will never again meet a ba-type suffix while continuing traversing the chain. The structure of ψ_{F_n} follows by Lemma 4(a). □

Theorem 7. *For $n \in \mathbb{N}$ even and $\beta \geq b$, let $B_n := \beta F_n$. $\mathsf{BWT}_{B_n} = b^{f_{n-2}} \beta a^{f_{n-1}}$ and $\psi_{B_n}[i] = f_n + 1$ if $i = f_{n-2} + 1$, $\psi_{B_n}[i] = 1$ if $i = f_n + 1$, $\psi_{B_n}[i] = (i + f_{n-1}) \mod f_n$ otherwise.*

Proof. The proof is conducted analogously to proof of Theorem 6: By Theorem 2, the smallest suffix is $B_n[f_n + 1..]$ and the step of SA_{B_n} is f_{n-2}; $B_n[f_n + 1..]$ is a ba-type suffix. By proof of Theorem 2, the largest suffix is $B_n[1..]$. By following the chain of the proof of Theorem 6, we traverse successively ba-type suffixes until we visit the ba-type suffix $\mathsf{SA}_{B_n}[f_{n-2}] = f_{n-1} + 2$. If $\beta \neq b$, we have already visited all suffixes of ba-type. Again, by a step of f_{n-2}, we find $\mathsf{SA}_{B_n}[f_{n-2} + 1] = 2$. The suffix $B_n[2..]$ is preceded by a β. □

Theorem 8. *For $n \geq 5$ odd, let $C_n := F_n c$. $\mathsf{BWT}_{C_n} = b^{f_{n-2}-1} c a^{f_{n-1}} b$ and $\psi_{C_n}[i] = f_n + 1$ if $i = f_{n-2}$, $\psi_{C_n}[i] = f_n$ if $i = f_n + 1$, $\psi_{C_n}[i] = (i + f_{n-1}) \mod f_n$ otherwise.*

Proof. The proof is conducted analogously to proof of Theorem 6: By Corollary 1(a), the smallest suffix is $C_n[f_{n-1}..]$ and the step of SA_{C_n} is f_{n-1}; $C_n[f_{n-1}..]$ is a ba-type suffix. By Theorem 3, the largest suffix is $C_n[f_n + 1..]$; it is preceded by a b character. Since $C_n[f_n..]$ is the second largest suffix, $\mathsf{BWT}_{C_n}[f_n + 1] = b$. By following the chain of the proof of Theorem 6, we traverse successively ba-type suffixes until we visit the ba-type suffix $\mathsf{SA}_{C_n}[f_{n-2} - 1] = f_{n-2} + 1$. Again, by a step of f_{n-1}, we find $\mathsf{SA}_{C_n}[f_{n-2}] = 1$. The suffix $C_n[1..]$ is preceded by a c. With Lemma 4(b) we yield the structure of $\psi_{F_n c}$. □

Theorem 9. *For $n \geq 4$ even, let $D_n := \bar{F}_n c$. $\mathsf{BWT}_{D_n} = b^{f_{n-1}-1} c a^{f_{n-2}} b$ and $\psi_{D_n}[i] = f_n + 1$ if $i = f_{n-1}$, $\psi_{D_n}[i] = f_n$ if $i = f_n + 1$, $\psi_{D_n}[i] = (i + f_{n-2}) \mod f_n$ otherwise.*

Proof. The proof is conducted by complementing the results of Theorem 6: Substrings of length two in D_n are of the form bc, bb, ba and ab; D_n does not contain the string aa. By Corollary 1(b), the smallest suffix is $D_n[f_{n-1}..]$ and the step of SA_{D_n} is f_{n-1}; $D_n[f_{n-1}..]$ is a ba-type suffix.

Because the suffixes starting with an a are always preceded by a b (suffixes of ba-type), it suffices to show that the suffixes starting with a b that are preceded by a b (suffixes of bb-type) are consecutively aligned after the block of ba-type suffixes. By Theorem 4, the largest suffix is $D_n[f_n + 1..]$; it is preceded by a b character. Since $D_n[f_n..]$ is the second largest suffix, $\mathsf{BWT}_{D_n}[f_n + 1] = b$. Moreover, the largest ba-type suffix is $D_n[f_n - 2..]$. By following a chain similar in the proof of Theorem 6, we traverse successively ba-type suffixes until we visit the ba-type suffix $\mathsf{SA}_{D_n}[f_{n-2}] = f_n - 1$. We next visit the bb-type suffixes starting from $\mathsf{SA}_{D_n}[f_{n-2} + 1] = f_{n-1} - 1$ to $\mathsf{SA}_{D_n}[f_{n-1} - 1] = f_{n-2} + 1$. By a step of f_{n-1}, we find $\mathsf{SA}_{D_n}[f_{n-1}] = 1$. The suffix $D_n[1..]$ is preceded by a c. □

Theorem 10. *For $n \geq 4$ even, $\mathsf{BWT}_{Z_n} = b^{f_{n-2}} a^{f_{n-1}-1} b$ and $\psi_{Z_n}[i] = (i + f_{n-1}) \mod f_n$, where Z_n is the n-th singular word, defined in Definition 3.*

Proof. Like F_n, the string Z_n does not contain bb as a substring. The proof is conducted analogously to proof of Theorem 6: By Lemma 9, the largest suffix is $Z_n[1..]$. It is preceded by $Z_n[|Z_n|] = b$; hence $\mathsf{BWT}_{Z_n}[f_n] = b$.

Moreover, the largest ba-type suffix is $Z_n[2..]$. By Theorem 5, the smallest suffix is $Z_n[f_{n-2}+1..]$ and the step of SA_{Z_n} is f_{n-2}; $Z_n[f_{n-2}..]$ is a ba-type suffix since $Z_n[f_{n-2}-1..f_{n-2}] = \delta_{n-2}$. By following the chain of the proof of Theorem 6, we traverse successively ba-type suffixes until we visit the last ba-type suffix $\mathsf{SA}_{Z_n}[f_{n-2}-1] = 2$. □

6 Outlook

We presented a family of string sequences based on the intriguing Fibonacci words, and highlighted some interesting combinatorial properties of the suffix array, its inverse, and the BWT of those sequences. It is an open question whether we can specify a general class of strings having the studied properties, like the BWT based on rotations [14]. One problem is that the conditions are not symmetric. Although we considered appending c or prepending β to a Fibonacci string, sequences like $\beta\bar{F_n}$, $\bar{F_n}\alpha$ and $F_n\alpha$ with $\alpha \le a$ do not show any nice properties. Even giving a characterization of the set of strings whose SA and ISA are identical seems hard for us. The here presented techniques could be useful for further research.

Acknowledgement We are grateful to Gabriele Fici for helpful discussion, and to our student Sven Schrinner who discovered one rotation property while solving an exercise.

References

[1] Berstel, J., Savelli, A.: Crochemore Factorization of Sturmian and Other Infinite Words. In: Královič, R., Urzyczyn, P. (eds.) Mathematical Foundations of Computer Science 2006, vol. 4162, pp. 157–166 (2006)

[2] Burrows, M., Wheeler, D.J., Burrows, M., Wheeler, D.J.: A block-sorting lossless data compression algorithm. Tech. rep., Digital Equipment Corporation (1994)

[3] Christodoulakis, M., Iliopoulos, C., Ardila, Y.: Simple Algorithm for Sorting the Fibonacci String Rotations. In: SOFSEM 2006, pp. 218–225 (2006)

[4] Droubay, X.: Palindromes in the Fibonacci word. Information Processing Letters 55(4), 217–221 (1995)

[5] Ferragina, P., Manzini, G.: Indexing compressed text. J. ACM 52(4), 552–581 (2005)

[6] Gagie, T., Gawrychowski, P., Kärkkäinen, J., Nekrich, Y., Puglisi, S.: LZ77-Based Self-indexing with Faster Pattern Matching. In: Pardo, A., Viola, A. (eds.) LATIN 2014, vol. 8392, pp. 731–742 (2014)

[7] Gramss, T.: Entropy of the symbolic sequence for critical circle maps. Phys. Rev. E 50, 2616–2620 (Oct 1994)

[8] Iliopoulos, C.S., Moore, D., Smyth, W.F.: A Characterization of the Squares in a Fibonacci String. Theor. Comput. Sci. 172(1-2), 281–291 (1997)

[9] Jr, V.E.H., Bicknell-Johnson, M.: Composites and Primes Among Powers of Fibonacci Numbers increased or decreased by one. Fibonacci Quarterly 15, 2 (1977)

[10] Kärkkäinen, J.: Fast bwt in small space by blockwise suffix sorting. Theor. Comput. Sci. 387(3), 249–257 (Nov 2007)

[11] de Luca, A.: A combinatorial property of the Fibonacci words. Information Processing Letters 12(4), 193–195 (1981)

[12] de Luca, A., de Luca, A.: Combinatorial Properties of Sturmian Palindromes. International Journal of Foundations of Computer Science 17(03), 557–573 (2006)

[13] Manber, U., Myers, G.: Suffix arrays: A new method for on-line string searches. In: Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 319–327. SODA '90, Philadelphia, PA, USA (1990)

[14] Mantaci, S., Restivo, A., Sciortino, M.: Burrows-Wheeler transform and Sturmian words. Inf. Process. Lett. 86(5), 241–246 (2003)

[15] Monnerot-Dumaine, A.: The Fibonacci Word fractal (Feb 2009), 24 pages, 25 figures

[16] Pirillo, G.: Fibonacci numbers and words. Discrete Mathematics 173(1-3), 197–207 (1997)

[17] Rytter, W.: The structure of subword graphs and suffix trees of Fibonacci words. Theor. Comput. Sci. 363(2), 211–223 (2006)

[18] Saari, K.: Periods of Factors of the Fibonacci Word. In: Proceedings of WORDS 2007. Institut de Mathematiques de Luminy (2007)

[19] Sadakane, K.: Compressed Suffix Trees with Full Functionality. Theory Comput. Syst. 41(4), 589–607 (2007)

[20] Simpson, J., Puglisi, S.J.: Words with Simple Burrows-Wheeler Transforms. Electr. J. Comb. 15(1) (2008)

[21] Wells, D.: Prime Numbers: The Most Mysterious Figures in Math. Wiley (2011)

[22] Wen, Z.X., Wen, Z.Y.: Some Properties of the Singular Words of the Fibonacci Word. Eur. J. Comb. 15(6), 587–598 (Nov 1994)

[23] Ziv, J., Lempel, A.: A universal algorithm for sequential data compression. IEEE Transactions on Information Theory 23(3), 337–343 (1977)