ELSEVIER

Theoretical Computer Science 218 (1999) 13–39

Theoretical Computer Science

# On the combinatorics of finite words

Aldo de Luca [a,b,*]

[a] *Dipartimento di Matematica, Università di Roma "La Sapienza", Piazzale A. Moro 2, 00185 Roma, Italy*

[b] *Istituto di Cibernetica del CNR, Arco Felice, Napoli, Italy*

**Abstract**

In this paper we consider a combinatorial method for the analysis of finite words recently introduced in Colosimo and de Luca (Special factors in biological strings, preprint 97/42, Dipt. Matematica, Univ. di Roma) for the study of biological macromolecules. The method is based on the analysis of (right) *special factors* of a given word. A factor $u$ of a word $w$ is special if there exist at least two occurrences of the factor $u$ in $w$ followed on the right by two distinct letters. We show that in the combinatorics of finite words two parameters play an essential role. The first, denoted by $R$, represents the minimal integer such that there do not exist special factors of $w$ of length $R$. The second, that we denote by $K$, is the minimal length of a factor of $w$ which cannot be extended on the right in a factor of $w$. Some new results are proved. In particular, a new characterization in terms of special factors and of $R$ and $K$ is given for the set $PER$ of all words $w$ having two periods $p$ and $q$ which are coprimes and such that $|w| = p + q - 2$.


*Keywords:* Special factors; Subword complexity; Sturmian words

## 1. Introduction

The study of the combinatorial properties of finite, as well as infinite, sequences of symbols over a finite set is a subject of great interest with remarkable applications in various fields such as Algebra, Physics, Computer Science and Biology. The set of symbols is usually called *alphabet* and the finite or infinite sequences *words* or *infinite words*, respectively.

As regards the applications, for instance, the existence of *unavoidable regularities* in very large words has many important consequences for the study of *finiteness conditions* for semigroups, groups and further algebraic structures (cf. [15]). Moreover, the *uniform recurrence* of infinite nonperiodic words such as Sturmian words and

* Correspondence address: Dipartimento di Matematica, Università di Roma "La Sapienza", Piazzale A. Moro 2, 00185 Roma, Italy E-mail: deluca@mercurio.mat.uniroma1.it.

0304-3975/99/$ - see front matter 
PII: S0304-3975(98)00248-5

Thue/Morse words, is an old subject of investigation with applications in Physics (for instance, the theory of *quasi-crystals* (cf. [5])).

The combinatorics of finite words is in some aspects a more recent subject of research. For instance, the study of repetitions and periodicities in finite words (as the theorem of *critical point* (cf. [18])) has a great interest in Computer Science for various questions related to 'Pattern Matching' and 'Data compression'. A further field which is a source of several combinatorial problems related to finite words, is molecular Biology. In fact, in this discipline the basic objects are biological sequences in a 4 letter alphabet (*DNA*, *RNA*) or in a 20 letter alphabet (*proteins*).

In a recent paper [7] devoted to this latter subject, we have introduced and developed a suitable technique of analysis of *DNA* sequences which allows us to obtain a large amount of information about the 'structure' of some 'genes'. The aim of this article is to consider in a general setting and with more details, this combinatorial analysis which is typical for finite words since it becomes trivial in the case of infinite words.

The basic notions of our analysis are the following. To each finite, or infinite word $w$ one can associate the language $F(w)$ of all finite factors, or finite blocks of consecutive letters, of $w$. A factor $u$ of $w$ is called *right special* (*left special*) if there exist two letters $x$ and $y$ of the alphabet such that $x \neq y$ and $ux,\ uy \in F(w)$, $(xu,\ yu \in F(w))$, i.e. there are two occurrences of $u$ in $w$ which are followed on the right (on the left) by two distinct letters. A factor of $w$ is called *bispecial* if it is right and left special. In the case of an alphabet with more than two letters one can have special factors of different *valence* and *order*. A right special factor has valence (order) $j \geqslant 2$ if it can be extended on the right in $w$ by $j$ (at least $j$) distinct letters.

Special and bispecial factors of infinite words have been studied by several authors [10–12, 14, 2, 6]. In the case of a finite word $w$ an important parameter which is always defined is the integer $R$ representing the least integer such that there are no right special factors of $w$ of length $\geqslant R$. In the case of an infinite word $R$ is a finite quantity if and only if the word is ultimately periodic. A further meaningful quantity is the integer $K$ defined as the minimal length of a factor of $w$ which cannot be extended in $w$ on the right by one letter. In the case of an infinite word (from left-to-right) this quantity is infinite.

We show that these two quantities $R$ and $K$ are two basic parameters in the combinatorial description of a finite word. The paper is organized as follows. In Section 3 we introduce the notions of *right-valence* of a factor of a word and the notion of special factor. In Section 4 the *subword complexity* $f_w$ of a finite or infinite word $w$ is considered. For any integer $n \geqslant 0$, $f_w(n)$ counts the number of distinct factors of $w$ of length $n$. A basic recursive formula which holds for any finite or infinite word is shown. This allows one to compute the subword complexity $f_w(n)$ for all $n > 0$ in terms of the distribution on the length of right (left) special factors of different valence (order). Moreover, in the case of finite words, some general results concerning the behaviour and the maximal values of $f_w$ are proved. In Section 5 we show that the *maximal repetition* in a finite nonempty word is given by $\max\{R, K\} - 1$. In Section 6 we prove an interesting and useful formula which allows us to evaluate the *total complexity* of a

word expressed in terms of the length of the word, the value of $K$ and the distribution of right special factors. In Section 7 we study the subword complexity of finite Sturmian words. We give a new characterization of the set *PER* of all words $w$ which have two periods $p$ and $q$ which are coprimes and such that $|w| = p+q-2$. Moreover, we prove that the total complexity of a word $w \in PER$ is given by $pq$. Finally, in Section 8 an equality relating subword complexity, the distribution of right special factors and the value of $K$ is proved by using the tree representation of a finite word.

## 2. Preliminaries

In this section we shall introduce some notations and definitions which will be used in the paper. In the following $\mathcal{A}$ will denote a finite *alphabet*, i.e. a finite nonempty set whose elements are called *letters*. By $\mathcal{A}^+$ we denote the set of all finite sequences of letters, or finite *words*. A finite word, or simply word, $w$ can be uniquely represented by a juxtaposition of its letters:

$$w = w_1 \dots w_n,$$

with $w_i \in \mathcal{A}$, $1 \leqslant i \leqslant n$. The integer $n$ is called the *length* of $w$ and is denoted by $|w|$. The set $\mathcal{A}^+$ of all the words over $\mathcal{A}$ is the *free semigroup* on $\mathcal{A}$, where the semigroup operation, called *product*, is defined by concatenation or juxtaposition of the words. If one adds to $\mathcal{A}^+$ the *identity* element $\varepsilon$, called *empty word*, then one obtains the *free monoid* $\mathcal{A}^*$ over $\mathcal{A}$. The length of $\varepsilon$ is taken to be equal to 0.

A word $u$ is a *factor*, or *subword*, of $w$ if there exist words $p, q \in \mathcal{A}^*$ such that $w = puq$. If $p$ ($q$) is equal to $\varepsilon$, then $u$ is called *prefix* (*suffix*) of $w$. We denote by $\mathrm{Pref}(w)$ ($\mathrm{Suf}(w)$) the set of all prefixes (suffixes) of $w$. For any pair $(i,j)$ of integers such that $1 \leqslant i \leqslant j \leqslant n$ we denote by $w[i,j]$ the factor $w[i,j] = w_i \dots w_j$.

If $w = w_1 \dots w_n$, $w_i \in \mathcal{A}$, $i = 1, \dots, n$, is a word, then the *reversed* $w^\sim$ of $w$ is the word

$$w^\sim = w_n \dots w_1.$$

Moreover, one sets $\varepsilon^\sim = \varepsilon$. A word is *palindrome* if $w^\sim = w$. The set of all palindromes will be denoted by *PAL*.

Let $w = w_1 \dots w_n$, $w_i \in \mathcal{A}$, $1 \leqslant i \leqslant n$, be a word. A positive integer $q$ is called a *period* of $w$ if $q \geqslant |w|$ or if $q < |w|$ the following condition is satisfied:

$$w_i = w_{i+q}$$

for all $i \in [1, n-q]$. We shall denote by $p_w$, or simply $p$, the *minimal period* of $w$. A word $w$ is called *periodic* if $p \leqslant [|w|/2]$, where for a real $x$, $[x]$ denotes its integer part.

In the following $\mathbb{N}$ ($\mathbb{N}_+$) will denote the set of nonnegative (positive) integers. An infinite (from left-to-right) word $w$ over the alphabet $\mathcal{A}$ is any map

$$w : \mathbb{N}_+ \to \mathcal{A}.$$

For each $n > 0$, we set $w_n = w(n)$ and denote $w$ also as

$$w = w_1 w_2 w_3 \dots .$$

A word $u \in \mathscr{A}^+$ is a finite *factor* of $w$ if there exist integers $i, j \in \mathbb{N}$, $0 < i \leqslant j$, such that $u = w_i \dots w_j$; the sequence $w[i,j] = w_i \dots w_j$ is also called an *occurrence* of $u$ in $w$.

An infinite word $w$ is called *ultimately periodic* if it can be expressed as

$$w = uv^\omega = uvv \dots v \dots ,$$

with $u \in A^*$ and $v \in A^+$.

For any finite or infinite word $w$, $F(w)$ denotes the set of all its finite factors and alph($w$) the set of all the letters of the alphabet $\mathscr{A}$ occurring in $w$, i.e. alph($w$) = $F(w) \cap \mathscr{A}$.

A *language* $L$ over the alphabet $\mathscr{A}$ is any subset of $\mathscr{A}^*$. For each $n \geqslant 0$ we denote by $\mathscr{A}^n$ the set of all the words of length $n$. For any language $L$ one denotes by $F(L)$ the set of the factors of all the words of $L$, i.e.

$$F(L) = \bigcup_{w \in L} F(w).$$

A language $L$ is *closed by factors* if $L = F(L)$. For any finite or infinite word $w$, $F(w)$ is, trivially, closed by factors.

## 3. Special factors

Let card($\mathscr{A}$) = $d$ and $w$ be a finite or infinite word over the alphabet $\mathscr{A}$. For any factor $u$ of $w$ we consider the maximal subset, with respect to the inclusion, $R_u$ of $\mathscr{A}$, that we simply denote by $R$, such that

$$uR \subseteq F(w).$$

Thus for all letters $x \in R$ one has $ux \in F(w)$ and, on the contrary, for all $x \in \mathscr{A} \setminus R$, $ux \notin F(w)$, i.e. $u$ occurs in $w$ followed on the right by any one of the letters of $R$, and only by these.

In a symmetric way one can consider the maximal subset $L_u$ of $\mathscr{A}$, that we simply denote by $L$, such that

$$Lu \subseteq F(w),$$

so that for all $x \in L$ , $xu \in F(w)$ and for all $x \in \mathscr{A} \setminus L$ one has $xu \notin F(w)$.

Let us now introduce the map $v_r : F(w) \to \mathbb{N}$ defined for all $u \in F(w)$ as

$$v_r(u) = \text{card}(uR) = \text{card}(R) = \text{card}(u\mathscr{A} \cap F(w)).$$

The integer $v_r(u)$ will be called the *right-valence* of $u$. In a symmetric way one can introduce the map $v_l : F(w) \to \mathbb{N}$ defined for all $u \in F(w)$ as

$$v_l(u) = \mathrm{card}(Lu) = \mathrm{card}(L) = \mathrm{card}(\mathscr{A}u \cap F(w)).$$

The integer $v_l(u)$ will be called the *left-valence* of $u$.

For any $u \in F(w)$ one has $0 \leqslant v_r(u), v_l(u) \leqslant d$. It easily follows that if a subword $u$ of $w$ has right (left) valence equal to $j$, then any suffix (prefix) $v$ of $u$ has right (left) valence $\geqslant j$. Let us observe that the empty subword $\varepsilon$ of $w$ has, according to the definition, a right and a left valence equal to $\mathrm{card}(\mathrm{alph}(w))$.

Let us give the following example. Let $w$ be the word of length 12

$$w = abcadabacada$$

on the four letter alphabet $\{a,b,c,d\}$. The subword $ab$ has right-valence equal to 2 since $ab$ occurs in $w$ followed on the right by the letter $a$ and by the letter $c$, whereas all the other factors of $w$ of length 2 have right-valence equal to 1. For instance, $ca$ occurs in $w$ followed on the right only by the letter $d$. The subword $a$ has right-valence equal to 3 since $a$ occurs in $w$ followed on the right by the letters $b$, $c$ and $d$. The subword $b$ has right valence 2 and the subwords $c$ and $d$ have right valence equal to 1. The subword $acada$ has right-valence equal to 0 since it does not occur in $w$ followed on the right by any letter. The left-valence of $cada$ is 2 since $cada$ can be followed on the left in $w$ by the letters $a$ and $b$. The left valence of $abc$ is 0.

A factor $u$ of $w$ is said of right (left) *order* $k$ if $v_r(u) \geqslant k$ ($v_l(u) \geqslant k$). Hence, if $u$ has order $k$ it can be extended on the right (left) in a factor of $w$ with at least $k$ distinct letters.

Let us observe that if $w$ is an infinite word, then any factor $u$ of $w$ can always be extended on the right in a factor of $w$ by at least one letter, so that $v_r(u) \geqslant 1$. However, the left-valence of some prefixes of the word can be 0. For instance, in the case of the word $ab^\omega = ab\ldots b\ldots$ the left-valence of all the prefixes $ab^j$ with $j \geqslant 0$, is 0.

A factor $u$ of $w$ is called *right special* if there exist at least two letters $a, b \in \mathscr{A}$, $a \neq b$ such that

$$ua, ub \in F(w),$$

i.e. the right-valence $v_r(u) > 1$. Thus a right special factor has a right-order equal to 2. In the case of an alphabet $\mathscr{A}$ having only two letters any right special factor is of course of valence two. In the case of *DNA* (*RNA*) alphabet having four letters one can have right special factors of valence 2, 3 and 4.

In a symmetric way one says that a factor $u$ of $w$ is *left special* if its left-valence $v_l(u) > 1$.

Since any suffix (prefix) $v$ of a right (left) special factor $u$ has a right (left) valence $v_r(v) \geqslant v_r(u)$ ($v_l(v) \geqslant v_l(u)$) one has that a suffix (prefix) of a right (left) special factor is still a right (left) special.

A factor $u$ of $w$ is said to be *bispecial* if it is both right and left special.

We shall refer to right special factors even though what we say can be extended, in a symmetric way, to the case of left special factors. In the following we shall denote by $S_r(w)$ the set of all right special factors of $w$. Moreover, for any $j$ such that $0 \leqslant j \leqslant d$ we denote by $S_r(j,w)$ the set of all factors of $w$ of valence $j$. One has, of course

$$S_r(w) = \bigcup_{1<j\leqslant d} S_r(j,w).$$

We set for $n \geqslant 0$,

$$s_r(j,n,w) = \mathrm{card}(S_r(j,w) \cap \mathscr{A}^n).$$

Let us, moreover, introduce for any $j$ such that $0 \leqslant j \leqslant d$ the following set:

$$G_r(j,w) = \bigcup_{j\leqslant k\leqslant d} S_r(k,w) = \{f \in F(w) \mid v_r(f) \geqslant j\},$$

which is the set of all factors of $w$ having a right-valence greater than or equal to $j$, i.e. of a right-order $j$. Moreover, we set for $n \geqslant 0$:

$$G_r(j,n,w) = G_r(j,w) \cap \mathscr{A}^n$$

and

$$g_r(j,n,w) = \mathrm{card}(G_r(j,n,w)) = \sum_{k=j}^{d} s_r(k,n,w).$$

Let us observe that for any $n \geqslant 0$ the following relation holds:

$$\sum_{j=2}^{d} (j-1) s_r(j,n,w) = \sum_{j=2}^{d} g_r(j,n,w). \tag{1}$$

In a symmetric way one can define the sets $S_l(w)$, $S_l(j,w)$, $G_l(j,w)$ and the maps $s_l(j,n,w)$ and $g_l(j,n,w)$. One has that $g_r(2,n,w)$ ($g_l(2,n,w)$) equals the number of right special (left special) factors of length $n$ of $w$. We shall simply denote $g_r(2,n,w)$ by $R_w(n)$ and $g_l(2,n,w)$ by $L_w(n)$.

In the following we shall drop in the formulas the reference to the word $w$ when there are no ambiguities.

## 4. Subword complexity

In this section we introduce the important notion of *subword complexity* of a finite or infinite word. Let $w$ be a word. The subword complexity $f_w$ of $w$ is the map $f_w : \mathbb{N} \to \mathbb{N}$ defined as

$$f_w(n) = \mathrm{card}(F(w) \cap \mathscr{A}^n).$$

For any $n$, $f_w(n)$ counts the number of distinct factors of length $n$ occurring in $w$. The subword complexity of infinite words has been extensively studied. A recent overview on this subject is in [1]. In this paper we shall be mainly concerned with finite words.

When $w$ is a finite word of length $N$, then $f_w(n) = 0$ for $n \geqslant N$. The quantity

$$c(w) = \sum_{i=0}^{N} f_w(i) = \mathrm{card}(F(w))$$

is called the *total complexity*, or *complexity index* of $w$. It gives a global measure of the 'richness' of the language of the subwords of $w$. The total complexity of a finite word has been considered, with different motivations, by several authors [17, 19, 7, 16]. In [17] a more general class of measures, called *d-complexities*, $d > 0$, was considered. For $d = 1$ one obtains the complexity index of $w$.

We denote by $C(w)$ the maximal value of the complexity $f_w(n)$ for $n \geqslant 0$, i.e.

$$C(w) = \max\{f_w(n) \mid n \geqslant 0\}.$$

We call $C(w)$ the *maximal complexity* of $w$.

When $w$ is an infinite (from left-to-right) word then any factor $u$ of $w$ can always be extended on the right by at least one letter in a factor of $w$. The situation is different in the case of a finite word $w$. Indeed, there can be subwords of $w$ which cannot be extended on the right in $F(w)$. Such words have to be, of course, suffixes of $w$.

We shall denote by $S_0(w)$ the set $S_r(0, w)$ of all factors of $w$ which cannot be extended on the right in $F(w)$, i.e. their right-valence is 0 and set for any $n \geqslant 0$

$$s_0(n) = \mathrm{card}(S_r(0, w) \cap \mathscr{A}^n) = s_r(0, n).$$

We shall set

$$K_w = \inf\{n \mid s_r(0, n) \neq 0\}.$$

If $w$ is an infinite word, then $K_w = \infty$. If $w$ is a finite word and $|w| = N$, then we denote by $k_w$ the suffix of $w$ of *minimal length* which cannot be extended on the right in $F(w)$ (Equivalently, $k_w$ is the suffix of $w$ of minimal length which has only one occurrence in $w$). One has that any word $\lambda k_w \in \mathrm{Suf}(w)$, $\lambda \in \mathscr{A}^*$, also cannot be extended on the right in $F(w)$.

Thus one has that $K_w = |k_w|$, so that $s_0(n) = 0$ for $0 \leqslant n \leqslant K_w - 1$ and $s_0(n) = 1$ for $K_w \leqslant n \leqslant N$. Hence, $s_0(n) \leqslant 1$, for all $0 \leqslant n \leqslant N$ and

$$\mathrm{card}(S_0(w)) = N - K_w + 1.$$

For any $w \in A^+$ the value of $K_w$ is such that $0 < K_w \leqslant N$. If $w = \varepsilon$, then we shall assume $K_\varepsilon = 0$. In the following we shall drop in $K_w$ the subscript $w$ when there is no ambiguity.

The following basic iterative equation holds: for $0 \leqslant n \leqslant N$,

$$f_w(n+1) = f_w(n) + \sum_{j=2}^{d} (j-1) s_r(j, n) - s_0(n). \tag{2}$$

The preceding equation relates the number of factors of $w$ of length $n+1$ with the number of factors of $w$ of length $n$ and the amounts of right special factors of length

$n$ having a different valence. Indeed, for each $n$ such that $0 \leqslant n \leqslant N$ there can exist at most one factor of length $n$ which has right-valence 0, so that it cannot be extended on the right in a factor of length $n+1$. Thus there will be at least $f_w(n) - s_0(n)$ factors of length $n+1$. If $u$ is a factor of $w$ of length $n$ having a valence $j > 1$, then $u$ can be extended on the right by further $j-1$ letters producing $j-1$ further subwords of length $n+1$.

Since $s_0(n) = s_r(0,n)$ the equation can be also rewritten in a more compact form as

$$f_w(n+1) = f_w(n) + \sum_{j=0}^{d} (j-1) s_r(j,n). \tag{3}$$

We note that the preceding equation holds true also when $n > N$ and for $n = 0$. In this latter case one has to recall that the empty word $\varepsilon$ is a right and left special factor of $w$ of valence equal to card(alph($w$)). By iteration of Eq. (3), since $f_w(0) = 1$, one obtains the following formula for the subword complexity of a finite, as well infinite word $w$ (cf. [7]):

$$f_w(n) = 1 + \sum_{k=0}^{n-1} \sum_{j=0}^{d} (j-1) s_r(j,k). \tag{4}$$

In the case of an infinite word $w$, one has $s_0(n) = 0$ for all $n \geqslant 0$, and Eq. (4) simply becomes

$$f_w(n) = 1 + \sum_{k=0}^{n-1} \sum_{j=2}^{d} (j-1) s_r(j,k). \tag{5}$$

Let us consider for any $n \geqslant 0$ the number $R_w(n)$ of all right special factors of $w$ of length $n$, i.e. $R_w(n) = g_r(2,n)$. Any suffix of a right special factor is still a right special factor; thus if there exists an integer $n$ for which $R_w(n) = 0$, then $R_w(m) = 0$ for all $m \geqslant n$. One can, define, $R_w$, or simply $R$, the quantity (possibly infinite if the word is infinite)

$$R = \inf\{n \mid R_w(n) = 0\}.$$

Thus if $0 < R < \infty$, then $R-1$ represents the *maximal length of a right special factor of* $w$. One has $R = 0$ if and only if the word $w$ has no right special factors. Thus also the empty word $\varepsilon$ is not special. This occurs if and only if the finite (infinite) word is a power ($\omega$-power) of a single letter. In the case of a finite word $w$ of length $N$ one has $R_w(N-1) = R_w(N) = 0$, so that $R_w$ is always defined. The following proposition shows that in the case of infinite words, with the only exception of ultimately periodic words, $R_w$ is always infinite.

**Proposition 4.1.** *An infinite word $w$ is ultimately periodic if and only if $R_w < \infty$.*

**Proof.** From a classic theorem on ultimately periodic words (cf. [8]) one has that an infinite word $w$ is ultimately periodic if and only if there exists an integer $n_0$ such that

$f_w(n_0) = f_w(n_0 + 1)$. From Eq. (3) this occurs, since $s_0(n) = 0$ for all $n \geqslant 0$, if and only if $R_w(n_0) = 0$.

Let us now suppose that $w$ is a finite word of length $N$. For $n = N$ one has $f_w(N) = 1$ so that from Eqs. (4) and (1), one derives, since $\sum_{n=0}^{N-1} s_r(0,n) = N-K$, that

$$\sum_{k=0}^{N-1} \sum_{j=2}^{d} (j-1)s_r(j,k) = \sum_{k=0}^{N-1} \sum_{j=2}^{d} g_r(j,k) = N - K. \tag{6}$$

Inverting the order of the sums in the preceding equation, one obtains [7]

$$\sum_{j=2}^{d} P_j = N - K \tag{7}$$

having set

$$P_j = \sum_{k=0}^{N-1} g_r(j,k). \tag{8}$$

For any $j \geqslant 2$, $P_j$ gives the *total number of right special factors of order j*. Since $P_2 = \mathrm{card}(S_r(w))$, then from Eq. (7) one derives the following upper bound to the total number of right special factors of $w$:

$$\mathrm{card}(S_r(w)) \leqslant N - K.$$

Let us now refer to left special factors of $w$. In this case one considers the prefix $h_w$ of minimal length which cannot be extended on the left in a subword of $w$. We shall denote by $H_w$, or simply $H$, the length of $h_w$, i.e. $H_w = |h_w|$. In a perfect symmetric way one can prove that for all $n \geqslant 0$ one has

$$f_w(n+1) = f_w(n) + \sum_{j=0}^{d} (j-1)s_l(j,n). \tag{9}$$

A formula similar to Eq. (7) can be derived for left special factors:

$$\sum_{j=2}^{d} \Lambda_j = N - H, \tag{10}$$

where for any $j \geqslant 2$, $\Lambda_j$ gives the *total number of left special factors of order j*. In a symmetric way by Eq. (10) one derives

$$\mathrm{card}(S_l(w)) \leqslant N - H.$$

From Eqs. (7) and (10) one obtains

$$\sum_{j=2}^{d} (P_j - \Lambda_j) = H - K.$$

If one compares Eqs. (3) and (9), then one derives the following important equation relating right and left special factors distributions: for all $n \geqslant 0$,

$$\sum_{j=0}^{d} (j-1)s_r(j,n) = \sum_{j=0}^{d} (j-1)s_l(j,n). \tag{11}$$

Let us set for all $n \geqslant 0$ and $x \in \{r, l\}$

$$\rho_x(n) = \sum_{j=2}^{d} (j-1) s_x(j,n) = \sum_{j=2}^{d} g_x(j,n). \tag{12}$$

We call $\rho_r$ ($\rho_l$) also the *weighted distribution* of the right (left) special factors of $w$. Let us define for all $n \geqslant 0$

$$\Delta_n = \rho_r(n) - \rho_l(n).$$

Since $s_r(0,n), s_l(0,n) \leqslant 1$ one derives from Eq. (11) that for all $n \geq 0$

$$|\Delta_n| \leqslant 1.$$

We shall now study the behaviour of the subword complexity $f_w$ of a finite word $w$ over $\mathscr{A}$ of length $N$. Let us first observe that the following upper bound to the subword complexity $f_w$ exists for all $0 \leqslant n \leqslant N$

$$f_w(n) \leqslant \min\{d^n, N - n + 1\}.$$

Indeed, $d^n$ is the cardinality of $\mathscr{A}^n$ and $N - n + 1$ is the set of all occurrences of subwords of length $n$ in $w$. The map $h : \mathbb{N} \to \mathbb{N}$ defined for $n \geqslant 0$ as

$$h(n) = \min\{d^n, N - n + 1\}$$

when $d > 1$ and $n$ is sufficiently small relatively to $N$, increases as an exponential with $n$ and decreases as a straight line having slope $-1$ (corresponding to an angle of $3\pi/4$). This passage from the exponential to the straight line occurs, for a value of $n$, that we denote by $e_N$, defined as

$$e_N = \min\{n \in \mathbb{N} \mid h(n) = N - n + 1\}.$$

One easily verifies that the following properties hold:
(a) $d^{e_N - 1} + e_N - 2 < N \leqslant d^{e_N} + e_N - 1$,
(b) $h$ takes its maximal value in $e_N$,
(c) $[\log_d N] \leqslant e_n \leqslant [\log_d N] + 1$.

Let us now get more information about $f_w$. As we shall see the values of $R$ and $K$ will play an essential role in the behaviour of $f_w$. The following proposition is in [7]. A similar proposition was proved independently by J. Cassaigne (private communication).

**Proposition 4.2.** *Let $w$ be a word of length $N$ such that* $\mathrm{card}(\mathrm{alph}(w)) > 1$ *and set* $m = \min\{R, K\}$ *and* $M = \max\{R, K\}$. *The subword complexity $f_w$ is strictly increasing in the interval $[0, m]$, is nondecreasing in the interval $[m, M]$ and strictly decreasing in the interval $[M, N]$. Moreover, for $n$ in the interval $[M, N]$, one has $f_w(n+1) = f_w(n) - 1$. If $R < K$, then $f_w$ is constant in the interval $[m, M]$.*

**Proof.** We distinguish two cases.

*Case* 1: $R<K$: For $n \in [0,R-1]$ one has that $s_0(n) = 0$ and $\rho_r(n) > 0$. Thus from Eq. (2), $f_w$ is *strictly increasing* in the interval $[0,R]$. For $n \in [R,K-1]$, $s_0(n) = 0$ and $\rho_r(n) = 0$, so that $f_w$ is *constant* in the interval $[R,K]$. For $n \in [K,N]$ one has $s_0(n) = 1$ and $\rho_r(n) = 0$, so that $f_w$ is *strictly decreasing* in the interval $[K,N]$, and, moreover, for $n \in [K,N]$

$$f_w(n+1) = f_w(n) - 1.$$

*Case* 2. $R \geqslant K$ : For $n \in [0,K-1]$ one has that $s_0(n) = 0$ and $\rho_r(n) > 0$, so that from Eq. (2), $f_w$ is *strictly increasing* in the interval $[0,K]$. For $n \in [K,R-1]$, $s_0(n) = 1$ and $\rho_r(n) > 0$, so that $f_w$ is *nondecreasing* in the interval $[K,R]$. For $n \in [R,N]$ one has $s_0(n) = 1$ and $\rho_r(n) = 0$, so that $f_w$ is *strictly decreasing* in the interval $[R,N]$, and, moreover, for $n \in [R,N]$

$$f_w(n+1) = f_w(n) - 1.$$

If one refers to left special factors, then one can define the quantity

$$L = \min\{n \mid L_w(n) = 0\}.$$

In a symmetric way one proves a dual proposition of Proposition 4.2, in which $R$ is replaced by $L$ and $K$ by $H$.

**Proposition 4.3.** *The subword complexity $f_w$ of a word $w$ of length $N$ takes its maximal value in $R$ and, moreover,*

$$f_w(R) = N - \max\{R,K\} + 1.$$

**Proof.** If $w$ is the power of a single letter, then the result is trivial. Indeed, one has $R = 0$, $K = N$ and $f_w(n) = 1$ for all $n \in [0,N]$. Let us then suppose that $\mathrm{card}(\mathrm{alph}(w)) > 1$. The subword complexity $f_w$ takes its maximal value in $R$. Indeed, from Proposition 4.2 if $R \geqslant K$, then $f_w$ takes its maximum value in $R$. On the contrary, if $R < K$, then $f_w$ takes its maximum value in $K$. However, in this case $f_w$ is constant in the interval $[R,K]$, so that $f_w(R) = f_w(K)$ and $f_w$ reaches in $R$ its maximum value. If $R \geqslant K$ one has that $f_w(n+1) = f_w(n) - 1$ for all $n$ in the interval $[R,N-1]$. This implies $1 = f_w(N) = f_w(R) - (N-R)$ and then $f_w(R) = N - R + 1$. If $R < K$ one has that $f_w(n+1) = f_w(n) - 1$ for all $n$ in the interval $[K,N-1]$. From this one derives $1 = f_w(N) = f_w(K) - (N-K)$. Since $f_w(K) = f_w(R)$ the result follows.

The maximal complexity $C(w)$ of a finite word is then given by

$$C(w) = N - \max\{R,K\} + 1. \tag{13}$$

Thus $C(w)$ depends only on the length $|w| = N$ of the word $w$ and on the values of $R$ and $K$.

By a symmetric argument one can prove the following:

**Proposition 4.4.** *The subword complexity $f_w$ of a word $w$ of length $N$ takes its maximal value in $L$ and, moreover,*

$$f_w(L) = N - \max\{L,H\} + 1.$$

From Propositions 4.3 and 4.4, one easily derives

**Corollary 4.1.** *The subword complexity $f_w$ is such that $f_w(R) = f_w(L)$. Moreover, one has*

$$\max\{R,K\} = \max\{L,H\}.$$

**Example.** Let $\mathcal{A} = \{a,b\}$. The word $w = abbbbbaababaaab$ is such that $R = L = 5$, $K = 4$ and $H = 3$. Moreover, $f_w(5) = 11$ and this is the maximal value of the subword complexity of $w$. The word $w = abaaaaaa$ is such that $K = L = 6$ and $H = R = 2$ and $f_w(2) = f_w(6) = 3$. In the case of word $w = aaabaaaba$ one has $K = H = 6 > R = L = 3$ and $f_w(3) = f_w(6) = 4$.

**Corollary 4.2.** *Let $w \in \mathcal{A}^+$ be a word of length $N$. Then*

$$\max\{R,K\} \geqslant [\log_d N].$$

**Proof.** Let $M = \max\{R,K\}$. From Propositions 4.3 and 4.4, one derives

$$f_w(M) = N - M + 1.$$

Moreover, $f_w(M) \leqslant h(M) \leqslant d^M$. Thus

$$d^M \geqslant N - M + 1.$$

This can occur only if $M \geqslant [\log_d N]$.

The following proposition [7], whose proof we omit, concerns the "structure" of the right (left) special factors of maximal length of a given word.

**Proposition 4.5.** *Let $w$ be a word and $u$ be a right (left) special factor of $w$ of maximal length. One has that $u$ is either a prefix (suffix) of $w$ or bispecial. If $R > H (L > K)$, then $u$ is bispecial.*

**Proposition 4.6.** *Let $w \in \mathcal{A}^*$ be a word of length $N$. Then*

$$N \geqslant R + K.$$

**Proof.** The result is trivial if $R = 0$ or $K = 0$. Indeed, in such a case $K = N$. Let us then suppose $R,K > 0$. We set $m = \min\{R,K\} > 0$. One has that for any integer $i$ such that $0 \leqslant i \leqslant m$ one has $f_w(i) \geqslant i+1$. Indeed, from Eq. (2) since for all $i \in [0, m-1]$, $s_0(i) = 0$ one has:

$$f_w(i) \geqslant f_w(i-1) + 1, \quad i = 1, \ldots, m,$$

so that since $f_w(0) = 1$, by iteration, the assertion follows.

Let us now first suppose $R \leqslant K$. One has then $m = R$ and then $f_w(R) \geqslant R + 1$. Moreover, from Proposition 4.3 we have that $f_w(R) = f_w(K) = N - K + 1$. From this one derives $N \geqslant R + K$. Let us now suppose $R > K$. In such a case $m = K$ and $f_w(K) \geqslant K + 1$. From Proposition 4.3 one has $f_w(K) \leqslant f_w(R) = N - R + 1$. From this it follows again $N \geqslant R + K$.

In a symmetric way one can prove that if $w \in \mathscr{A}^*$ is a word of length $N$, Then

$$N \geqslant L + H. \tag{14}$$

**Proposition 4.7.** *Let $w$ be a word of length $N$, $m = \min\{R,K\}$ and $M = \max\{R,K\}$. One has that $N = R + K$ if and only if*

$$f_w(i) = i + 1 \quad \textit{for} \quad i = 0,1,\ldots,m,$$

$$f_w(i+1) = f_w(i) \quad \textit{for} \quad i = m,\ldots,M-1,$$

$$f_w(i+1) = f_w(i) - 1 \quad \textit{for} \quad i = M,\ldots,N.$$

*Moreover, $f_w(R) = f_w(K)$ and the maximal value of $f_w$ is $m + 1$.*

**Proof.** The result is trivial if $\mathrm{card}(\mathrm{alph}(w)) \leqslant 1$. Let us then suppose that $\mathrm{card}(\mathrm{alph}(w)) > 1$. We first suppose that $R \leqslant K$ so that $m = R$ and $M = K$. In such a case we know that $f_w$ is strictly increasing in the interval $[0,R]$, $f_w$ is constant in the interval $[R,K]$ and strictly decreasing in the interval $[K,N]$; moreover,

$$f_w(n+1) = f_w(n) - 1, \quad \text{for} \quad n \in [K,N]. \tag{15}$$

Hence, we have to prove only that $f_w(i) = i + 1$, for $i = 0,1,\ldots,R$. Since $f_w(0) = 1$, one has then $f_w(i) \geqslant i + 1$ for $i = 0,1,\ldots,R$. Thus $f_w(R) \geqslant R + 1$. Let us now prove that $f_w(R) = R + 1$. Indeed, from Eq. (15) and the hypothesis $R = N - K$ one derives

$$1 = f_w(N) = f_w(K) - N + K = f_w(K) - R.$$

Since $f_w(K) = f_w(R)$ one obtains $f_w(R) = R + 1$. This implies that $f_w(i) = i + 1$, for $i = 0,1,\ldots,R$. Indeed, if by contradiction, there exists $i_0 < R$ for which $f_w(i_0) > i_0 + 1$, then from the basic recursive formula (cf. Eq. (2)) $f_w(i_0 + k) \geqslant f_w(i_0) + k$ with $k = 1,\ldots,R - i_0$. Hence, $f_w(R) = f_w(i_0) + R - i_0 > R + 1$ which is absurd.

Let us now consider the case $K < R$, so that $m = K$ and $M = R$. In this case $f_w$ is strictly increasing in the interval $[0,K]$, nondecreasing in the interval $[K,R]$ and strictly decreasing in the interval $[R,N]$; moreover, for $n \in [R,N]$

$$f_w(n+1) = f_w(n) - 1.$$

The maximal value of $f_w$ is reached in $R$, having, since $N = R + K$,

$$f_w(R) = N - R + 1 = K + 1.$$

Since $K < R$ one has $f_w(K) \geqslant K+1$, so that $f_w(R) = f_w(K) = K+1$. By using an argument similar to that used in the preceding case it follows that

$$f_w(i) = i+1 \quad \text{for} \quad i = 0,\ldots,K.$$

Conversely, suppose first that $K < R$. By hypothesis $f_w(i) = i+1$ for $i = 0,\ldots,K$, so that $f_w(K) = K+1$. Moreover, since $f_w(i+1) = f_w(i)$ for $i \in [K, R-1]$ it follows $f_w(K) = f_w(R)$. Finally, by the condition $f_w(i+1) = f_w(i) - 1$ for $i \in [R,N]$, one derives $f_w(R) = N-R+1$. Hence, $K+1 = N-R+1$, i.e. $N = K+R$. Let us now suppose $K \geqslant R$. One easily derives from the hypothesis that $f_w(R) = R+1 = f_w(K) = N-K+1$ that implies $N = K+R$.

By a symmetric argument one can prove a dual proposition of Proposition 4.7, in which $R$ is replaced by $L$ and $K$ by $H$.

The following proposition shows that if a word $w$ of length $N$ is such that $N = R+K$, then $N = L+H$.

**Proposition 4.8.** *Let $w$ be a word of length $N$. If $N = R+K$, then*

$$\min\{R,K\} = \min\{L,H\}$$

*and $N = L+H$.*

**Proof.** The result is trivial if $\text{card}(\text{alph}(w)) \leqslant 1$. Let us then suppose that $\text{card}(\text{alph}(w)) > 1$. Let us first show that

$$\min\{L,H\} \leqslant \min\{R,K\}. \tag{16}$$

Indeed, by Corollary 4.1 and Eq. (14) one has

$$N = R+K = \min\{R,K\} + \max\{R,K\} = \min\{R,K\} + \max\{L,H\} \geqslant L+H,$$

so that $\min\{L,H\} \leqslant \min\{R,K\}$.

Let us now suppose that $L \leqslant H$. One has, by Corollary 4.1, $H = \max\{R,K\}$ and from Eq. (16), $L \leqslant \min\{R,K\}$. Since $L \leqslant H$ from Proposition 4.7 and the dual of Proposition 4.2, one has that $f_w(L) = f_w(H) = f_w(R) = f_w(K) = \min\{R,K\}+1$. This implies that $\min\{R,K\} \leqslant L \leqslant \max\{R,K\}$. Hence, $L = \min\{L,H\} = \min\{R,K\}$.

Let us suppose that $H \leqslant L = \max\{R,K\}$, so that by Eq. (16) one has $H \leqslant \min\{R,K\}$. Let us set in the following $m = \min\{R,K\}$. By Proposition 4.7 for any $i \in [0,m]$ one has

$$f_w(i) = i+1. \tag{17}$$

This implies that $\text{card}(\text{alph}(w)) = 2$ so that we set $\text{alph}(w) = \{a,b\}$. Moreover, for any $n \in [0, m-1]$ there is only one right special factor of length $n$ and for any $n \in [0, H-1]$ there is only one left special factor of length $n$. We shall now prove that for any $n \in [H-1, m-1]$ there is a unique left special factor of length $n$. The proof is by induction on the value of $n$. For $n = H-1$ the statement is true. Let us

then prove it for $H-1 < n \leqslant m-1$. Suppose then that there exist two left special factors $u$ and $v$ such that $u \neq v$ and $|u| = |v| = n$. This implies that

$$au,\ bu,\ av,\ bv \in F(w). \tag{18}$$

Let us write $u = u'x$, $v = v'y$ with $x, y \in \{a,b\}$. Since $u'$ and $v'$ are left special factors of $w$ of length $n-1$, then by the inductive hypothesis $u' = v' = s$. Moreover, since $u \neq v$, then $x \neq y$. Thus from Eq. (18) one has

$$asx,\ bsx,\ asy,\ bsy \in F(w).$$

This shows that $as$ and $bs$ are two right special factors of $w$ of length $n$ which is a contradiction. Hence, for any $n \in [H-1, m-1]$ there is a unique left special factor of length $n$. This implies by Eq. (17) that $f_w(m) = f_w(m-1) = m$ which is a contradiction. Hence, $H = \min\{R,K\}$. The remaining part of the proof is trivial.

**Proposition 4.9.** *The maximal complexity $C(w)$ of a word $w$ satisfies the inequality*

$$C(w) \geqslant \min\{R,K\} + 1.$$

*The lower bound is reached if and only if $|w| = N = R + K$.*

**Proof.** We know (cf. Eq. (13)) that $C(w) = N - \max\{R,K\} + 1$. From Proposition 4.6, $N \geqslant R + K$; hence $C(w) \geqslant R + K - \max\{R,K\} + 1$. Now $R + K - \max\{R,K\} = \min\{R,K\}$, so that $C(w) \geqslant \min\{R,K\} + 1$. Moreover, $C(w) = \min\{R,K\} + 1$ if and only if $N - \max\{R,K\} = \min\{R,K\}$, i.e. $N = R + K$.

## 5. Repetitions in a word

Let $w$ be a word on the alphabet $\mathscr{A}$. In $w$ there is a *repetition* of a factor $u$ of $w$ if in $w$ there are two distinct occurrences of $u$. More formally, $w$ has a repetition of the factor $u$ if

$$w \in \mathscr{A}^*(u\mathscr{A}^+ \cap \mathscr{A}^+u)\mathscr{A}^*,$$

i.e. there exist words $\lambda, \mu \in \mathscr{A}^*$ and $\alpha, \beta \in \mathscr{A}^+$ such that

$$w = \lambda f \mu \quad \text{with } f = u\alpha = \beta u.$$

Note that the two occurrences of the repeated factor $u$ can, in general, overlap. If they do not overlap, then we can simply write

$$w \in \mathscr{A}^*u\mathscr{A}^*u\mathscr{A}^*.$$

We say that in $w$ there is a repetition of *length $k$* if there is a repetition of a factor $u$ of $w$ such that $|u| = k$. To each word $w$ one can associate the quantity $r(w)$ defined as the *maximal length of a repetition of a factor of $w$*, i.e.

$$r(w) = \max\{|u| \mid w \in \mathscr{A}^*(u\mathscr{A}^+ \cap \mathscr{A}^+u)\mathscr{A}^*\}.$$

Note that this quantity is always defined since, obviously, $r(w) < N$, where $N$ is the length of $w$. If $w = a^N$, with $a \in \mathscr{A}$, then $r(w) = N - 1$.

**Theorem 5.1.** *Let $w \in \mathscr{A}^+$. One has*

$$r(w) = \max\{R, K\} - 1.$$

**Proof.** If $R = 0$, then there are no special factors of $w$, so that $w \in a^*$ with $a \in \mathscr{A}$. This implies $K = N$ and $r(w) = N - 1$. Let us then suppose that $R > 0$. Let $u$ be a right special factor of maximal length, i.e. $|u| = R - 1$. Then $u$ has to occur in $w$ followed by two distinct letters. Hence, $r(w) \geqslant R - 1$. Moreover, we know that the suffix $s$ of $w$ of length $K - 1$ has to occur in $w$ followed, on the right, by one letter of $\mathscr{A}$. Hence, $r(w) \geqslant K - 1$. Hence,

$$r(w) \geqslant \max\{R, K\} - 1. \tag{19}$$

Let us now prove the inverse inequality. We first suppose that $K < R$. Let $u$ be a repeating factor of maximal length, i.e. $|u| = r(w)$. We prove that there exist two letter $x, y \in \mathscr{A}$ such that $ux, uy \in F(w)$. Suppose, by contradiction, that this is not the case. This can occur only if $u$ is a suffix of the word and, moreover, $|u| < K < R$. This contradicts the fact that, by Eq. (19), $r(w) \geqslant R - 1$. Thus $ux, uy \in F(w)$. If $x = y$, then one would contradict the maximality of $|u|$. Let us assume that $x \neq y$. The factor $u$ is then right special so that

$$r(w) = |u| \leqslant R - 1.$$

Let us now suppose $R \leqslant K$. Let $u$ be a factor of $w$ such that there is a repetition of $u$ and $|u| = r(w) > K - 1$. The factor $u$ cannot be obviously a suffix of $w$. Thus there exist there exist two letter $x, y \in \mathscr{A}$ such that $ux, uy \in F(w)$. If $x = y$, then one contradicts the maximality of the length of $u$. Hence, $x \neq y$ and $u$ is a right special of length $|u| > K - 1 \geqslant R - 1$ which is a contradiction. Hence, $r(w) \leqslant \max\{R, K\} - 1$, which concludes the proof.

**Corollary 5.1.** *Let $w \in \mathscr{A}^+$ be a word of length $N$. One has*

$$C(w) = N - r(w).$$

**Proof.** Trivial from Eq. (13) and Theorem 5.1.

**Corollary 5.2.** *Let $w \in \mathscr{A}^+$ be a word of length $N$. Then*

$$r(w) \geqslant [\log_d N] - 1.$$

**Proof.** From Corollary 4.2 we know that $\max\{R, K\} \geqslant [\log_d N]$. Thus from Theorem 5.1, $r(w) = \max\{R, K\} - 1 \geqslant [\log_d N] - 1$.

**Proposition 5.1.** *Let $w$ be a word of length $N$ and $p$ be its minimal period. One has*

$$p \geqslant N - K + 1.$$

**Proof.** Let $u$ be the suffix of $w$ of length $K$. One can write

$$w = \lambda u \quad \text{with} \quad \lambda \in \mathcal{A}^*.$$

The subword $u$ cannot have any other occurrence in $w$. Since $w$ has the period $p$ and the first letter of $u$ is in the position $N - K + 1$, the preceding condition implies that

$$N - K + 1 - p < 1, \tag{20}$$

or $N - K < p$.

**Corollary 5.3.** *Let $w$ be a word of length $N$ and $p$ its minimal period. One has*

$$p \geqslant R + 1.$$

*If $p = R + 1$, then $N = R + K$.*

**Proof.** From the preceding proposition one has $p \geqslant N - K + 1$. By Proposition 4.6, $R \leqslant N - K$, so that $p \geqslant R + 1$. If we make the hypothesis that $p = R + 1$, then by Proposition 5.1 it follows $N \leqslant R + K$, so that from Proposition 4.6 the result follows.

**Example.** Let $\mathcal{A} = \{a, b, c\}$. The word $w = ababacbcabacba$ has $K = 3$ and $R = 6$. One has $r(w) = 5$. Indeed, $w$ has the right special factor $abacb$ and this is the factor of $w$ of maximal length which has a repeated occurrence in $w$. The word $w = aacbcabcacbca$ has $K = 6$ and $R = 3$. Hence, $r(w) = 5$. The factor $acbca$, which is not right special, occurs once inside the word and another time as suffix of $w$.

**Corollary 5.4.** *Let $w$ be a periodic finite word of length $N$ and minimal period $p$. One has*

$$R \leqslant [N/2] - 1, \; K \geqslant [N/2] + 1,$$

*and then $K \geqslant R + 2$. Moreover $p \leqslant K - 1$.*

**Proof.** Let $w$ be a periodic finite word having minimal period $p$. Thus $p \leqslant [N/2]$. From Corollary 5.3 one has $p \geqslant R + 1$ and then $R \leqslant [N/2] - 1$. Moreover, from Proposition 5.1, $p \geqslant N - K + 1$, so that $K \geqslant N - [N/2] + 1 \geqslant [N/2] + 1$. Hence, $K \geqslant R + 2$. Since $p \leqslant [N/2]$ there is in $w$ a repetition of length $p$. Thus $p \leqslant \max\{R, K\} - 1 = K - 1$.

Let us observe that, by a symmetric argument, one can replace in Proposition 5.1 and Corollaries 5.3 and 5.4, $R$ with $L$ and $K$ with $H$. As a consequence one derives that if $w$ is a word having length $N$ and minimal period $p$ then

$$p \geqslant N - \min\{K, H\} + 1.$$

This implies also that $p \geqslant \max\{R, L\} + 1$.

## 6. Complexity index of a finite word

In this section we shall give a simple formula which allows us to compute the complexity index of a finite word in terms of the length $N$ of the word, the value of $K$ and the distribution of right special factors. More precisely, let us define for each $j = 2,\ldots,d$ the quantity

$$\Omega_j = \sum_{n=0}^{R} n g_r(j,n). \tag{21}$$

One has:

**Theorem 6.1.** *Let $w$ be a word of length $N$ and $d = \mathrm{card}(\mathscr{A})$. One has*

$$c(w) = 1 + \frac{(N+K)(N-K+1)}{2} - \sum_{j=2}^{d} \Omega_j.$$

**Proof.** If $R = 0$, then $w = a^N$, $a \in \mathscr{A}$, and $K = N$. Moreover, $\Omega_j = 0$, $j = 2,\ldots,d$, so that $c(w) = 1 + N$. Let us then suppose $R > 0$. We recall that

$$c(w) = \sum_{i=0}^{N} f_w(i).$$

We can decompose this sum in two parts

$$c(w) = \sum_{i=0}^{R} f_w(i) + \sum_{i=R+1}^{N} f_w(i).$$

Let us compute the first term. From Eqs. (4) and (1) we can write

$$\sum_{i=0}^{R} f_w(i) = R + 1 + \sum_{n=0}^{R-1} \sum_{h=0}^{n} \sum_{j=2}^{d} g_r(j,h) - \sum_{n=0}^{R-1} \sum_{h=0}^{n} s_r(0,h).$$

Let us set

$$Z = \sum_{n=0}^{R-1} \sum_{h=0}^{n} \sum_{j=2}^{d} g_r(j,h).$$

One has

$$Z = \sum_{n=0}^{R-1} (R-n) \sum_{j=2}^{d} g_r(j,n) = R \sum_{n=0}^{R-1} \sum_{j=2}^{d} g_r(j,n) - \sum_{n=0}^{R-1} \sum_{j=2}^{d} n g_r(j,n).$$

By Eqs. (21) and (8) one derives

$$Z = R \sum_{j=2}^{d} P_j - \sum_{j=2}^{d} \Omega_j.$$

By Eq. (7) one has

$$Z = R(N-K) - \sum_{j=2}^{d} \Omega_j.$$

We can then write

$$c(w) = 1 + R(N - K + 1) - \sum_{j=2}^{d} \Omega_j + \sum_{i=R+1}^{N} f_w(i) - \sum_{n=0}^{R-1} \sum_{h=0}^{n} s_r(0,h).$$

In order to prove our result we have to distinguish two cases:

*Case* 1. $R \geqslant K$: Let us evaluate $\sum_{i=R+1}^{N} f_w(i)$. We recall that, by Proposition 4.2, $f_w$ is strictly decreasing in the interval $[R,N]$, and, moreover, for $n \in [R,N]$

$$f_w(n+1) = f_w(n) - 1.$$

Moreover, by Proposition 4.3, $f_w(R) = N - R + 1$. Hence,

$$\sum_{i=R+1}^{N} f_w(i) = \sum_{j=1}^{N-R} j = \frac{(N-R)(N-R+1)}{2}.$$

Let us now evaluate $\sum_{n=0}^{R-1} \sum_{h=0}^{n} s_r(0,h)$. One has

$$\sum_{n=0}^{R-1} \sum_{h=0}^{n} s_r(0,h) = \sum_{n=K}^{R-1} \sum_{h=K}^{n} 1 = \sum_{n=K}^{R-1} (n - K + 1) = \frac{(R-K)(R-K+1)}{2}.$$

By a simple calculation one derives that

$$R(N-K+1) + \frac{(N-R)(N-R+1)}{2} - \frac{(R-K)(R-K+1)}{2} = \frac{(N+K)(N-K+1)}{2},$$

so that in this case the result is proved.

*Case* 2. $K \geqslant R$. In this case we have that $f_w$ is *constant* in the interval $[R,K]$, *strictly decreasing* in the interval $[K,N]$, and, moreover, for $n \in [K,N-1]$

$$f_w(n+1) = f_w(n) - 1.$$

Hence, $f_w(R) = f_w(K) = N - K + 1$. Let us compute first $\sum_{i=R+1}^{N} f_w(i)$. One has

$$\sum_{i=R+1}^{N} f_w(i) = \sum_{i=R+1}^{K} f_w(i) + \sum_{i=K+1}^{N} f_w(i).$$

In the interval $[R+1,K]$, $f_w(i)$ takes the constant value $N - K + 1$, so that

$$\sum_{i=R+1}^{K} f_w(i) = (K-R)(N-K+1).$$

Moreover,

$$\sum_{i=K+1}^{N} f_w(i) = \sum_{j=1}^{N-K} j = \frac{(N-K)(N-K+1)}{2}.$$

Hence,

$$\sum_{i=R+1}^{N} f_w(i) = \frac{(N-K+1)(N+K-2R)}{2}.$$

In this case, since $R \leqslant K$, one has

$$\sum_{n=0}^{R-1} \sum_{h=0}^{n} s_r(0,h) = 0.$$

By a simple calculation one derives:

$$R(N-K+1) + \frac{(N-K+1)(N+K-2R)}{2} = \frac{(N-K+1)(N+K)}{2},$$

which proves our assertion.

**Proposition 6.1.** *Let $w$ be a word of length $N$. If $N = R+K$, then the complexity index of $w$ is*

$$c(w) = (R+1)(K+1).$$

**Proof.** The result is trivial if $R = 0$ or $K = 0$. Indeed, in such a case $w = a^N$, $a \in \mathcal{A}$, and $K = N$, so that $c(w) = N+1$. Let us then suppose $R, K > 0$. By Proposition 4.7 one has that $f_w(i) = i+1$ for $i = 0, 1, \ldots, m$ where $m = \min\{R, K\}$. We prove that for any length $n$ in the interval $[0, R-1]$ there is only one right special factor whose right-valence is 2. The result is, trivially, true if $K \geqslant R$. Let us then suppose that $K < R$. Our assertion is, obviously, true for $n \in [1, K-1]$. For $n \in [K, R-1]$ is a consequence of the fact that $f_w(n) = f_w(K) = f_w(R)$ and $s_r(0,n) = 1$ in the interval $[K, R-1]$.

Hence, in our case one has

$$\sum_{j \geqslant 2}^{d} \Omega_j = \Omega_2 = \sum_{n=1}^{R-1} n = \frac{R(R-1)}{2}.$$

From Theorem 6.1 and the fact that $N = R+K$, one derives

$$c(w) = (R+1)(K+1).$$

## 7. Finite Sturmian words

Sturmian words are infinite words $w$ whose subword complexity $f_w$ is such that

$$f_w(n) = n+1$$

for all $n \geqslant 0$, so that they have the minimal possible value for subword complexity without being ultimately periodic (cf. [9, 3]). Moreover, since $f_w(1) = 2$ one has that these words are in a two-letter alphabet. It is worth noting that between ultimately periodic and Sturmian words there are no other words. We shall denote by $St$ the set of the factors of all Sturmian words.

Let us observe that the condition $f_w(n) = n+1$ for any $n \geqslant 0$ is equivalent to the statement that for any length $n \geqslant 0$ the word $w$ has exactly one right special factor.

**Proposition 7.1.** *Let $w \in St$ be a word of length $N$. Then*

$$N = R + K.$$

**Proof.** Let $w \in St$ and $m = \min\{R,K\}$ and $M = \max\{R,K\}$. One has that for $i \in [0,m]$, $f_w(i) \geqslant i+1$. However, since $w$ is a factor of a Sturmian word then $f_w(i) \leqslant i+1$. Thus $f_w(i) = i+1$ for $i \in [0,m]$. Moreover, in the interval $[m,M]$ one has $f_w(i+1) = f_w(i)$. This is obvious if $R < K$. In the case $R \geqslant K$ one has to observe that for any $n$ in the interval $[K,R-1]$ there is one factor of length $n$ which cannot be prolonged on the right and one and only one right special factor of length $n$. Indeed, $n < R$ and $w$ is a factor of an infinite Sturmian word. Hence, in any case $f_w(R) = f_w(K)$. From the general behaviour of the function $f_w$ we know that $f_w$ is strictly decreasing in the interval $[M,N]$, and, moreover, for $n \in [M,N]$, $f_w(n+1) = f_w(n) - 1$. Thus by Proposition 4.7 the result follows.

The following example shows that the above condition, i.e. $N = R + K$ does not characterize finite Sturmian words.

**Example.** Consider the word $w = aaabab$ of length $N = 6$ which is not Sturmian. One has $R = K = 3$ so that $N = R + K$. The word $w = aaabbbb$ of length $N = 7$ is not Sturmian and such that $R = 3$ and $K = 4$.

Let us now consider the set *PER* of all words $w$ having two periods $p,q$ such that $\gcd(p,q) = 1$ and $|w| = p+q-2$. Thus, a word $w$ belongs to *PER* if it is a power of a single letter or is a word of maximal length for which the theorem of Fine and Wilf (cf. [18]) does not apply. In the sequel, we assume that $\varepsilon \in PER$. This is, formally, coherent with the above definition if one takes $p = q = 1$. The importance of the set *PER* for Sturmian words is due to the following result (cf. [9]):

$$St = F(PER),$$

i.e. the set of all finite factors of all infinite Sturmian words coincides with the set of all factors of the set *PER*. The set *PER* has several characterizations based on quite different concepts (cf. [10, 3, 9]). We mention here the following [10]. Let *Stand* be the set of all finite standard Sturmian words. One has

**Theorem 7.1.**

$$Stand = \{a,b\} \cup PER\{ab,ba\}.$$

We recall the following important structure result on the set *PER* whose proof is in [9].

**Proposition 7.2.** *Let $w$ be a word such that* $\mathrm{card}(\mathrm{alph}(w)) > 1$. *Then $w \in PER$ if and only if $w$ can be uniquely represented as*

$$w = PxyQ = QyxP,$$

*with $x, y$ fixed letters in $\{a,b\}$, $x \neq y$ and $P,Q \in PAL$. Moreover,* $\gcd(p,q) = 1$, *where $p = |P| + 2$ and $q = |Q| + 2$ are periods of $w$.*

**Proposition 7.3.** *Let $w \in PER$ and $N = |w|$. One has*

$$R = p - 1 \text{ and } K = q - 1,$$

*where $p$ is the minimal period of $w$ and $q = N + 2 - p$. Moreover, if $p > 1$ then the prefix (suffix) of $w$ of length $p - 2$ is the unique right (left) special factor of $w$ of maximal length.*

**Proof.** Let us first suppose that $\text{card}(\text{alph}(w)) \leqslant 1$, i.e. $w = a^N$. In such a case the result is trivially true since one has $p = 1$, $q = N + 1$, $K = N$ and $R = 0$ (note that in this case the empty word $\varepsilon$ is not a special factor). Let us then suppose that $\text{card}(\text{alph}(w)) > 1$. From Proposition 7.2 one has

$$w = PxyQ = QyxP, \tag{22}$$

with $x, y$ fixed letters in $\{a,b\}$, $x \neq y$ and $P,Q \in PAL$ and $|P| < |Q|$. From Proposition 5.1 one has $p \geqslant R + 1$. We shall now prove that $P$ is a right (left) special factor of $w$. Since $|Q| \geqslant |P| + 1$ one has from Eq. (22), $Q = Px\zeta$, $\zeta \in \{a,b\}^*$, and

$$w = PxyPx\zeta = \zeta^{\sim}xPyxP.$$

Thus $Px, Py \in F(w)$ so that $P$ is a right special factor. Hence, $|P| \leqslant R - 1$. Since $|P| = p - 2$ it follows $p \leqslant R + 1$. Thus, by Corollary 5.3, we have proved that $p = R + 1$. From Corollary 5.3, or from Proposition 7.1, one has $N = R + K$. By the fact that $N = p + q - 2$ it follows $K = q - 1$. From Proposition 4.7 one has that $P$ is the unique right (left) special factor of maximal length.

**Example.** Consider the word $w = abaababaaba \in PER$ of length 11. One has that the right special factor of maximal length is the prefix $aba$ so that $R = 4$. Moreover, the suffix of minimal length which cannot be extended on the right is $babaaba$, so that $K = 7$. In this case, as one easily verifies, the minimal period $p = 5$ and $q = 8$. Let us consider the word $w = aababaabaab \in St$ of length 11; one easily derives that $R = 5$, $K = 6$ and $p = 8$. Note that $w \notin PER$.

Let us recall the following lemma whose proof is in [9]:

**Lemma 7.1.** *A palindrome word $w$ has the period $p < |w|$ if and only if it has a palindrome prefix (suffix) of length $|w| - p$.*

**Theorem 7.2.** *Let $w$ be a word of length $N$ having a minimal period $p > 1$. Then $w \in PER$ if and only if*

(i) $w \in PAL$,

(ii) $K > R$,

(iii) *The prefix (suffix) of $w$ of length $p-2$ is a right (left) special factor of $w$ of maximal length.*

**Proof.** The 'only if' part of the theorem is obviously derived from Proposition 7.3. Let us then prove the 'if' part. By hypothesis $p-2=R-1$, so that $p=R+1$. By Corollary 5.3 one has $N=R+K$. This also implies, since $R>0$, by Proposition 4.7 that $f_w(1)=2$, so that $w$ is a word in a two letter alphabet $\mathscr{A}=\{a,b\}$. By Lemma 7.1 the word $w$ has a palindrome suffix $Q$ of length

$$|Q|=N-p=R+K-p=K-1,$$

so that $Q$ is the suffix (prefix) of maximal length which can be extended on the right (left) in $w$. Let us denote by $P$ the prefix of $w$ of length $p-2$. We can write, since $w$ is palindrome,

$$w=PxyQ=QyxP^{\sim},$$

with $x,y\in\{a,b\}$. Let us now prove that $P$ is palindrome. Indeed, $P$ is a right special factor of maximal length, so that there exists a letter $z\in\mathscr{A}$ such that $z\neq x$ and $Pz\in F(w)$. Since $K\geqslant R+1$ one has

$$N=K+R=K+p-1\geqslant 2p-1.$$

From the $p$-periodicity this implies that $w$ has the prefix $PxyPx$ whose length is $2p-1$. Hence, $yPx\in F(w)$. Moreover $Pz$ can be extended on the left in $w$. Indeed, since $w\in PAL$ then $H=K$, so that $|Pz|=R<H$. Thus there exists a letter $y'$ such that $y'Pz\in F(w)$. One has that $y\neq y'$ otherwise one would have a right special factor of $w$ of length $>R-1$ which is a contradiction. Since $w\in PAL$ then $xP^{\sim}y, zP^{\sim}y'\in F(w)$. This implies that $P^{\sim}$ is a right special factor of $w$. By Proposition 4.7 there can be only one right special factor of length $|P|$ so that $P=P^{\sim}$.

Let us now prove that $x\neq y$. Let us suppose, by contradiction, that $x=y$. Then $w$ will have the prefix $PxxPx$. Moreover, $Pz$ can be extended on the left in $w$ only by the letter $z\neq x$. Thus $zPz\in F(w)$. The word $zP$ cannot be a prefix of $w$. Indeed, otherwise one would have $zP=Px$ which is a contradiction since $P$ is palindrome. Thus $zPz$ can be extended on the left in $w$. Due to the $p$-periodicity one has $zzPz\in F(w)$ and then $zzP\in F(w)$. Thus there exists a prefix $f$ of $w$ such that

$$f=Pxx\lambda=\mu zzP.$$

From this equation one has from the lemma of Lyndon and Schützenberger (cf. [18]):

$$\mu zz=\alpha\beta,\ xx\lambda=\beta\alpha,\ P=(\alpha\beta)^n\alpha,$$

with $n\geqslant 0$ and $\alpha,\beta\in\mathscr{A}^*$. Since $P$ is a palindrome one derives $\alpha,\beta\in PAL$. Hence, one has

$$zz\mu^{\sim}=\beta\alpha=xx\lambda,$$

that implies $x = z$ which is a contradiction. Hence, $x \neq y$. We can write

$$w = PxyQ = QyxP.$$

From Proposition 7.2 the result follows.

The following example shows that in the 'if' part of the preceding theorem the condition (iii) cannot be replaced with the weaker requirement $p - 2 = R - 1$.

**Example.** The palindrome word $w = babaababaabab$ has $R = 4$, $K = 9$, $|w| = R+K = 13$ and minimal period $p = 5$. The word $w$ has the unique right special factor $aba$ of length 3, which is not a prefix of $w$. The word $w \notin PER$.

**Corollary 7.1.** *Let $w \in PER$ and $N$ be its length. One has*

$$c(w) = pq \quad \text{and} \quad C(w) = p,$$

*where $p$ is the minimal period of $w$ and $q = N + 2 - p$.*

**Proof.** Let $w \in PER$ and $N = |w|$. From Proposition 7.3 one has $R = p - 1$ and $K = q - 1$ and $N = R + K$. By Proposition 6.1 the complexity index of $w$ is $c(w) = (R + 1)(K + 1) = pq$. By Propositions 7.1 and 4.9, $C(w) = \min\{R, K\} + 1$, so that from Proposition 7.3, $C(w) = p$.

**Example.** The word $w = abaababaaba \in PER$ of length 11 has $p = 5$ and $q = 8$. One has $f_w(i) = i + 1$ for $0 \leqslant i \leqslant 3$, $f_w(i) = 5$ for $4 \leqslant i \leqslant 7$ and $f_w(i) = 12 - i$ for $8 \leqslant i \leqslant 11$. Thus $c(w) = pq = 40$ and $C(w) = 5$.

## 8. Tree representation

Let $\mathcal{A}$ be an alphabet of cardinality $d$. To each finite word $w \in \mathcal{A}^*$ one can associate a finite labeled tree $T_w$ as follows. One starts with the $d$-ary general tree $\mathcal{T}_d$. As is well known, there exists a one-to-one correspondence between the nodes of this tree and the words in a $d$ letter alphabet. We shall consider the case where the branches in the tree represent the covering relation of the prefixial ordering. The tree $T_w$ is obtained from $\mathcal{T}_d$ by taking all the nodes which represent factors of the word $w$. Thus, any factor of $w$ will be represented by a node $v$ of $T_w$ or, equivalently, by a unique path going from the root (representing the empty word) to $v$. The leaves of $T_w$ represent factors of $w$ which cannot be extended on the right in $w$, i.e. elements of the set

$$G_w = \mathcal{A}^* k_w \cap \mathrm{Suf}(w),$$

where $k_w$ is the suffix of $w$ of minimal length ($= K_w$) which cannot be extended on the right in $w$. The tree $T_w$ has an height equal to the length $N$ of the word $w$ and it is not complete. In order to complete $T_w$ one has to add to $T_w$ a certain number

of nodes corresponding to a set $F_w$ of words defined as follows. A word $u \in F_w$ if $u$ is not a factor of $w$, but the prefix of $u$ of length $|u|-1$ is a factor of $w$. Thus the completed tree $T_w^c$ will have a set of leaves represented by the set

$$X_w = F_w \cup G_w.$$

This set $X_w$ is a prefix maximal code (cf. [4]) so that if we set for all $n \geqslant 0$, $\phi_w(n) = \mathrm{card}(X_w \cap \mathscr{A}^n)$, then the Kraft–McMillan equality has to be satisfied:

$$\sum_{n \geqslant 0} \phi_w(n) d^{-n} = 1. \tag{23}$$

In the following we shall suppose that $\mathrm{card}(\mathrm{alph}(w)) = d > 1$.

**Proposition 8.1.** *Let $w$ be a word of length $N$. Then the subword complexity $f_w$ and the weighted distribution $\rho_r$ of the right special factors of $w$ satisfy the equality*

$$\sum_{n=0}^{N-1} ((d-1) f_w(n) - \rho_r(n)) d^{-(n+1)} = 1 - \frac{1}{d-1}(d^{-K} + (d-2)d^{-N}).$$

**Proof.** Let us observe that for any $n \geqslant 0$ the following recursive formula holds:

$$f_w(n+1) = d(f_w(n) - s_0(n)) - \phi_w(n+1) + s_0(n+1). \tag{24}$$

One has only to observe that $f_w(n)$ gives the number of all nodes at the height $n$ in the tree $T_w$. The number of nodes at the height $n+1$ in the complete tree $T_w^c$ is then $d(f_w(n) - s_0(n))$. In order to obtain $f_w(n+1)$ one has to subtract the quantity $\phi_w(n+1) - s_0(n+1)$ from the preceding number. Moreover, from Eqs. (2) and (12) one has

$$f_w(n+1) = f_w(n) + \rho_r(n) - s_0(n). \tag{25}$$

From Eqs. (24) and (25) one derives

$$(d-1) f_w(n) - \rho_r(n) = (d-1) s_0(n) + \phi_w(n+1) - s_0(n+1).$$

Thus, by using Eq. (23), it follows

$$\begin{aligned}
&\sum_{n=0}^{N-1} ((d-1) f_w(n) - \rho_r(n)) d^{-(n+1)} \\
&\quad = 1 + (d-1) \sum_{n=0}^{N-1} s_0(n) d^{-(n+1)} - \sum_{n=0}^{N-1} s_0(n+1) d^{-(n+1)} \\
&\quad = 1 + (d-1) \sum_{n=K}^{N-1} d^{-(n+1)} - \sum_{n=K}^{N} d^{-n}.
\end{aligned}$$

Since $(d-1)\sum_{n=K}^{N-1} d^{-(n+1)} - \sum_{n=K}^{N} d^{-n} = -(1/(d-1))(d^{-K} + (d-2)d^{-N})$, the result follows.

In the case of a word $w$ in a two letter alphabet (i.e. $d = 2$) the preceding proposition simply becomes:

**Corollary 8.1.** *Let $w$ be a word of length $N$ over a two-letter alphabet. Then the subword complexity $f_w$ and the distribution $R_w$ of right special factors satisfy the equality*

$$\sum_{n=0}^{N-1} (f_w(n) - R_w(n))2^{-(n+1)} = 1 - 2^{-K}.$$

**Proposition 8.2.** *The following relation holds*:

$$\mathrm{card}(X_w) = (d-1)(c(w) - 1 - N + K) + 1.$$

**Proof.** From Eq. (24) one has

$$\sum_{n=0}^{N-1} f_w(n+1) = d \sum_{n=0}^{N-1} f_w(n) - d \sum_{n=0}^{N-1} s_0(n) - \sum_{n=0}^{N-1} \phi_w(n+1) + \sum_{n=0}^{N-1} s_0(n+1).$$

Hence,

$$\begin{aligned} c(w) - 1 &= d(c(w) - 1)) - \mathrm{card}(X_w) - d \sum_{n=K}^{N-1} s_0(n) + \sum_{n=K}^{N} s_0(n) \\ &= d(c(w) - 1) - \mathrm{card}(X_w) - d(N-K) + N - K + 1. \end{aligned}$$

From this the result follows.

## References

[1] J.-P. Allouche, Sur la complexité des suites infinies, Bull. Belg. Math. Soc. 1 (1994) 133–143.
[2] M. Béal, F. Mignosi and A. Restivo, Minimal forbidden words and symbolic dynamics, Proc. STACS '96,
Lecture Notes in Computer Science, vol. 1046, Springer, Berlin, 1996, pp. 555–566.
[3] J. Berstel, A. de Luca, Sturmian words, Lyndon words and trees, Theoret. Comput. Sci. 178 (1997) 171–203.
[4] J. Berstel, D. Perrin, Theory of Codes, Academic Press, New York, 1985.
[5] E. Bombieri, J.E. Taylor, Which distributions of matter diffract? An initial investigation, J. Physique, Colloque C3 (suppl. 7), 47 (1986) 19–28.
[6] J. Cassaigne, Complexité et facteurs speciaux, Bull. Belg. Math. Soc. 4 (1997) 67–88.
[7] A. Colosimo, A. de Luca, Special factors in biological strings, preprint 97/42, Dipartimento di Matematica, Università di Roma 'La Sapienza'.
[8] E.M. Coven, G. Hedlund, Sequences with minimal block growth, Math. Systems Theory 7 (1973) 138–153.
[9] A. de Luca, Sturmian words: structure, combinatorics, and their arithmetics, Theoret. Comput. Sci. (special issue on Formal Languages) 183 (1997) 45–82.
[10] A. de Luca, F. Mignosi, Some combinatorial properties of Sturmian words, Theoret. Comput. Sci. 136 (1994) 361–385.
[11] A. de Luca and L. Mione, On bispecial factors of the Thue-morse word, Inform. Process. Lett. 49 (1994) 361–385.
[12] A. de Luca, S. Varricchio, On the factors of the Thue-morse word on three symbols, Inform. Process. Letters 27 (1988) 281–285.

[13] A. de Luca, S. Varricchio, Some combinatorial problems of the Thue-morse sequence and a problem in semigroups, Theoret. Comput. Sci. 63 (1989) 333–348.
[14] A. de Luca, S. Varricchio, A combinatorial theorem on $p$-power-free words and an application to semigroups, R,A,I,R,O,I,T 24 (1990) 205–228.
[15] A. de Luca, S. Varricchio, Regularity and finiteness conditions, chapter in: G. Rozenberg, A. Salomaa (Eds.), the Handbook on 'Formal Languages' vol. 1, Springer, Berlin, 1997, pp. 747–810.
[16] S. Ferenczi, Z. Kása, Complexities for finite factors of infinite sequences, preprint, 1998.
[17] A. Iványi, On the d-complexity of words, Annales Univ. Sci. Budapest. Sect. Comp. 8 (1987) 69–90.
[18] M. Lothaire, Combinatorics on Words, Addison-Wesley, Reading, MA, 1983.
[19] J. Shallit, On the maximum number of distinct factors of a binary string, Graphs Combin. 9 (1993) 197–200.