Counting maximal-exponent factors in words*

Golnaz Badkobeh, Maxime Crochemore,
and Robert Mercaş

Mathematics Subject Classification (2000) : 68R15.

1 Introduction

The topic of repeating segments in words is one of major interest in combinatorics on words. The topic has been studied for more than a century by many authors after the seminal work [9] which described infinite words containing no consecutive occurrences of the same factor.

Beyond the theoretical aspect of questions related to redundancies in words, repetitions, also called repeats in the following, are often the base for string modelling adapted to compression coding. They play an important role in run-length compression and in Ziv–Lempel compression, e.g., [4]. Moreover, repetitions receive considerable attention in connection with the analysis of genetic sequences. Their occurrences are called tandem repeats, satellites or SRS and accept some notion of approximation. The existence of some palindromic repeats is crucial for the prediction of the secondary structure of RNA molecules influencing their biological functions, see [5].

Repetitions are composed of consecutive occurrences of the same factor. Their occurrences have been extended to runs [8], maximal periodic factors, and their number has been shown to be less than the word length n [3] (see also [6]) and even further less than $22n/23$ [7].

In this article we consider factors that repeat non consecutively in a given word of length n. They are of the form uvu, where u is their longest border (factor occurring both at the beginning and end of the word). Their exponent, defined as the ratio of their length over their smallest period length, that is, $|uvu|/|uv|$, is smaller than 2. The number of occurrences of these factors may be quadratic with respect to the word length even if they are restricted to non extensible occurrences. This is why we focus on factors having the maximal exponent among all factors occurring in a square-free word. They are called maximal-exponent factors, MEFs in short, and thus have all the same exponent.

*This represents an extended abstract of a work accepted for publication in *Theor. Comput. Sci.*, doi:10.1016/j.tcs.2016.02.035

The first attempt to count the number of occurrences of MEFs is done in [2]. In there, authors restrict themselves to considering square-free words, and prove that this number is upper bounded by $2.25n$. They also give an example of a word containing $0.66n$ such factors. The reason for restricting the question to only square-free words, words that contain no factor with an exponent at least 2, comes from the question related to the maximum number of runs in a word. If the word contains squares, the maximal exponent of factors is at least 2 and MEF occurrences become runs whose largest number is known to be less than the word length (see [3, 6, 7]).

2 Results

Considering the interactions between MEFs with borders that are not double one each other, and have a quite long overlap, we prove the following result.

Theorem 2.1 *There are less than $4n/b$ occurrences of MEFs with maximum length border at least b in a length n word.*

As a direct consequence of Theorem 2.1, one can count the number of MEFs with border length at least b for any positive b. We choose $b = 8$ in this paper because of the way we structured the counting of all MEFs.

Corollary 2.1 *There are less than $n/2$ occurrences of MEFs with border length at least 8 in a word of length n.*

Next, we look at the positioning of overlaps between two MEFs, one of which has border length twice of the other.

Proposition 2.1 *There are at most $2n/(2\ell + 1)$ MEFs with border lengths ℓ and 2ℓ in a word of length n.*

Following Proposition 2.1 there are at most $2n/3$ MEFs with border lengths 1 and 2, $2n/5$ MEFs with borders 2 and 4, $2n/7$ MEFs with borders 3 and 6, and so on. Next lemma is a further refinement on the number of MEFs with border lengths that are small and exponentially increasing.

Lemma 2.1 *Every word of length n contains at most $4n/5$ MEFs with the border length in the set $\{1, 2, 4\}$.*

Next result is a summation of the previous ones and represents the final stride towards improving the upper bound on the number of MEFs.

Theorem 2.2 *There exist at most $1.8n$ number of occurrences of MEFs in a word of length n.*

Although the upper bound in Theorem 2.2 is true in general, this bound can be further improved when special cases of MEFs are considered.

Lemma 2.2 *There are at most n occurrences of maximal-exponent factors in a word of length n, whenever the maximal exponent is greater than* 1.5.

Theorem 2.3 *Every length n word contains at most n occurrences of MEFs whenever the length of these factors is not a multiple of their longest border.*

We end with an example of a construction that generates a word that has a ratio of 5/6 of MEF occurrences relative to its length with the maximal exponent 10/9. This improves the result presented in [1, Section 6.2].

In the following we consider the fixed alphabet

$$\Sigma = \{\mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3, \mathtt{b}_1, \mathtt{b}_2, \mathtt{b}_3, \mathtt{b}_4, \mathtt{c}_1, \mathtt{c}_2, \mathtt{c}_3, \mathtt{c}_4, \mathtt{d}_1, \mathtt{d}_2, \mathtt{d}_3, \mathtt{d}_4, \mathtt{e}_1, \mathtt{e}_2, \mathtt{e}_3, \mathtt{e}_4\},$$

and the infinite alphabet

$$\Sigma_\infty = \{\mathtt{f}_{1,1}, \mathtt{f}_{2,1}, \ldots, \mathtt{f}_{8,1}, \mathtt{f}_{1,2}, \mathtt{f}_{2,2}, \ldots\}.$$

We define the following sequence for $i > 0$:

$$u_{(1,i)} = \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_1\ \mathtt{a}_2\ \mathtt{d}_1\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_1\ \mathtt{f}_{1,i} \qquad u_{(2,i)} = \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_2\ \mathtt{a}_2\ \mathtt{d}_2\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_1\ \mathtt{f}_{2,i}$$
$$u_{(3,i)} = \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_3\ \mathtt{a}_2\ \mathtt{d}_3\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_2\ \mathtt{f}_{3,i} \qquad u_{(4,i)} = \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_4\ \mathtt{a}_2\ \mathtt{d}_4\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_2\ \mathtt{f}_{4,i}$$
$$u_{(5,i)} = \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_1\ \mathtt{a}_2\ \mathtt{d}_5\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_3\ \mathtt{f}_{5,i} \qquad u_{(6,i)} = \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_2\ \mathtt{a}_2\ \mathtt{d}_6\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_3\ \mathtt{f}_{6,i}$$
$$u_{(7,i)} = \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_3\ \mathtt{a}_2\ \mathtt{d}_7\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_4\ \mathtt{f}_{7,i} \qquad u_{(8,i)} = \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_4\ \mathtt{a}_2\ \mathtt{d}_8\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_4\ \mathtt{f}_{8,i}$$

and the infinite word $\Omega = \prod_{i=1}^{\infty} \left(\prod_{j=1}^{8} u_{(j,i)} \right)$.

Proposition 2.2 *The ratio between the length of the prefixes of Ω and the number of occurrences of the MEFs they contain tends to* 5/6.

Note that the maximal exponent of factors in Ω is 10/9 and that its construction can be extended to whatever exponent of the form $(2^\ell+2)/(2^\ell+1)$, in a similar fashion. It is also our belief that this construction can be generalised as to generate, for any integer ℓ, an infinite word Ω_ℓ in which every MEF has a border length of the form 2^i, $i \leq \ell$, and whose asymptotic number of MEF occurrences per position grows very closely to 1 with ℓ.

Finally, observe that letters $\mathtt{f}_{j,i}$ occurring in Ω can be drawn from an 11-letter alphabet disjoint from Σ. To do so, it suffices to replace the infinite subsequence of $\mathtt{f}_{j,i}$ by an infinite sequence whose maximal exponent of factors is 11/10, Dejean's repetitive threshold of the alphabet. No MEFs considered in the previous proof will be affected.

References

[1] G. Badkobeh, M. Crochemore, Computing maximal-exponent factors in an overlap-free word, *J. Comput. Syst. Sci.*, **82(3)** (2016), 477–487.

[2] G. Badkobeh, M. Crochemore, C. Toopsuwan, Computing the maximal-exponent repeats of an overlap-free string in linear time, *19th SPIRE*, **7608 LNCS** (2012), 61–72.

[3] H. Bannai, T. I, S. Inenaga, Y. Nakashima, M. Takeda, K. Tsuruta, The "runs" theorem, *CoRR*, **abs/1406.0263** (2014).

[4] T. C. Bell, J. G. Clearly, I. H. Witten, Text Compression, *Prentice Hall Inc.*, New Jersey, 1990.

[5] H.-J. Böckenhauer, D. Bongartz, Algorithmic Aspects of Bioinformatics, *Springer*, Berlin, 2007.

[6] M. Crochemore, R. Mercaş, On the density of Lyndon roots in factors, *Theor. Comput. Sci.*, **in press** (2016).

[7] J. Fischer, Š. Holub, T. I, M. Lewenstein, Beyond the runs theorem, *22nd SPIRE*, **9309 LNCS** (2015), 277–286.

[8] C. S. Iliopoulos, D. Moore, W. F. Smyth, A characterization of the squares in a Fibonacci string, *Theor. Comput. Sci.*, **172(1–2)** (1997), 281–291.

[9] A. Thue, Über unendliche Zeichenreihen, *Norske vid. Selsk. Skr. I. Mat. Nat. Kl. Christiana*, **7** (1906), 1–22.

Golnaz Badkobeh
Department of Computer Science, University of Warwick, UK
E-mail: `Golnaz.Badkobeh@gmail.com`

Maxime Crochemore
Department of Informatics, King's College London, UK
E-mail: `Maxime.Crochemore@kcl.ac.uk`

Robert Mercaş
Department of Computer Science, Loughborough University, UK
E-mail: `robertmercas@gmail.com`