Discovering Hidden Repetitions in Words

Paweł Gawrychowski[1], Florin Manea[2], Dirk Nowotka[2]

[1] Max-Planck-Institut für Informatik, Saarbrücken, Germany,
gawry@cs.uni.wroc.pl
[2] Christian-Albrechts-Universität zu Kiel, Institut für Informatik, Kiel, Germany,
{flm,dn}@informatik.uni-kiel.de

Abstract. Pseudo-repetitions are a natural generalization of the classical notion of repetitions in sequences: they are the repeated concatenation of a word and its encoding under a certain morphism or antimorphism. We approach the problem of deciding whether there exists an anti-/morphism for which a word is a pseudo-repetition. In other words, we try to discover whether a word has a hidden repetitive structure. We show that some variants of this problem are efficiently solvable, while some others are NP-complete. This manuscript is an abstract of [3].
Keywords: repetition, pseudo-repetition, pattern matching, stringology.

1 Definitions

Let V be a finite alphabet. We denote by V^* the set of all words over V and by V^k the set of all words of length k. The *length* of a word $w \in V^*$ is denoted by $|w|$. The *empty word* is denoted by λ. Moreover, we denote by $\mathrm{alph}(w)$ the alphabet of all letters that occur in w. In the problems discussed in this paper we are given as input a word w of length n and we assume that the letters of w are in fact integers from $\{1, \dots, n\}$ and w is seen as a sequence of integers. This is a common assumption in algorithmic on words (see, e.g., [4]).

A word u is a *factor* of a word v if $v = xuy$, for some x, y; also, u is a *prefix* of v if $x = \lambda$ and a *suffix* of v if $y = \lambda$. We denote by $w[i]$ the symbol at position i in w and by $w[i..j]$ the factor $w[i]w[i+1] \dots w[j]$ of w starting at position i and ending at position j. For simplicity, we assume that $w[i..j] = \lambda$ if $i > j$. A word u occurs in w at position i if u is a prefix of $w[i..|w|]$. The powers of a word w are defined recursively by $w^0 = \lambda$ and $w^n = ww^{n-1}$ for $n \geq 1$. If w cannot be expressed as a power of another word, then w is *primitive.* If $w = u^n$ with $n \geq 2$ and u primitive, then u is called the primitive root of w. A *period* of a word w over V is a positive integer p such that $w[i] = w[j]$ for all i and j with $i \equiv j \pmod p$. By $per(w)$ we denote the smallest period of w.

A function $f : V^* \to V^*$ is a morphism if $f(xy) = f(x)f(y)$ for all $x, y \in V^*$; f is an antimorphism if $f(xy) = f(y)f(x)$ for all $x, y \in V^*$. Note that to define an anti-/morphism it is enough to give the definitions of $f(a)$, for all $a \in V$. We say that f is *uniform* if there exists a number k with $f(a) \in V^k$, for all $a \in V$; if $k = 1$ then f is called *literal.* If $f(a) = \lambda$ for some $a \in V$, then f is called *erasing*, otherwise *non-erasing.* The vector T_f of $|V|$ natural numbers with

$T_f[a] = |f(a)|$ is called the length-type of the anti-/morphism f in the following. If $V = \{a_1, \ldots, a_n\}$, T is a vector of n natural numbers $T[a_1], \ldots, T[a_n]$, and $x = b_1 \cdots b_k$ with $b_i \in V$ for all i, we denote by $T(x) = \sum_{i \le k} T[b_i]$, the length of the image of x under any anti-/morphism of length type T defined on V.

We say that a word w is an *f-repetition*, or, alternatively, an f-power, if w is in $t\{t, f(t)\}^+$, for some prefix t of w; for simplicity, if $w \in t\{t, f(t)\}^+$ then w is called an f-power of root t. If w is not an f-power, then w is *f-primitive*.

For example, the word *abcaab* is primitive from the classical point of view (i.e., **1**-primitive, where **1** is the identical morphism) as well as f-primitive, for the morphism f defined by $f(a) = b$, $f(b) = a$ and $f(c) = c$. However, when considering the morphism $f(a) = c$, $f(b) = a$ and $f(c) = b$, we get that *abcaab* is the concatenation of ab, $ca = f(ab)$, and ab, thus, being an f-repetition.

2 Overview

In [2], an efficient solution for the problem of deciding, given a word w and an anti-/morphism f, whether w is an f-repetition was given. Here we approach a more challenging problem. Namely, we are interested in deciding whether there exists an anti-/morphism f for which a given word w is an f-repetition. Basically, we check whether a given word has an intrinsic (yet hidden) repetitive structure. Note that in the case approached in [2] the main difficulty was to find a prefix x of w such that $w \in x\{x, f(x)\}^*$. The case we discuss here seems more involved: not only we need to find two factors x and y such that $w \in x\{x, y\}^*$, i.e., a suitable decompositions of w, but we also have to decide the existence of an anti-/morphism f with $f(x) = y$. The problem is defined in the following.

Problem 1. Given $w \in V^+$, decide whether there exists an anti-/morphism $f : V^* \to V^*$ and a prefix t of w such that $w \in t\{t, f(t)\}^+$.

The unrestricted version of the problem is, however, trivial. We can always give a positive answer for input words of length greater than 2. It is enough to take the (non-erasing) anti-/morphism f that maps the first letter of w, namely $w[1]$, to $w[2..n]$, where $n = |w|$. Clearly, $w = w[1]f(w[1])$, so w is indeed an f-repetition. When the input word has length 1 or 0, the answer is negative.

On the other hand, when we add a series of simple restrictions to the initial statement, the problem becomes more interesting. The restrictions we define are of two types: either we restrict the desired form of f, and try to find anti-/morphisms of given length type, or we restrict the repetitive structure of w by requiring that it consists in at least three repeating factors or that the root of the pseudo-repetition has length at least 2.

In the first case, when the input consists both in the word w and the length type of the anti-/morphism we are trying to find, we obtain a series of polynomial time solutions for Problem 1. More precisely, in the most general case we can decide whether there exists an anti-/morphism f such that w is an f-repetition in $\mathcal{O}(n(\log n)^2)$ time. Note that deciding whether a word is an f-repetition when f is known took only $\mathcal{O}(n \log n)$ time [2]. When we search for an uniform morphism we solve the problem in optimal linear time time. This matches the complexity of deciding, for a given uniform anti-/morphism f, whether a given word is an f-repetition, obtained in [2]. This result covers also the case of literal anti-/morphism, extensively approached in the literature (see, e.g., [1, 5]). Our solutions are based both on combinatorial results regarding the structure of pseudo-repetitions and on the usage of efficient data-structures.

For the second kind of restrictions, the length type of f is no longer given. In this case, we want to check, for instance, whether there exist a prefix t and an anti-/morphism f such that w is an f-repetition that consists in the concatenation of at least 3 factors t or $f(t)$. The most general case as well as the case when we add the supplementary restriction that f is non-erasing are NP-complete; the case when f is uniform (but of unknown length type) is tractable. The problem of checking whether there exists a prefix t, with $|t| \geq 2$, and a non-erasing anti-/morphism f such that $w \in t\{t, f(t)\}^+$ is also NP-complete; this problem becomes tractable for erasing or uniform anti-/morphisms.

Our two main theorems are:

Theorem 2. *Given a word w and a vector T of $|V|$ numbers, we decide whether there exists an anti-/morphism f of length type T such that $w \in t\{t, f(t)\}^+$ in $\mathcal{O}(n(\log n)^2)$ time. If T defines uniform anti-/morphisms we need $\mathcal{O}(n)$ time.*

Theorem 3. *For a word $w \in V^+$, deciding the existence of an anti-/morphism $f : V^* \to V^*$ and a prefix t of w such that $w \in t\{t, f(t)\}^+$ with $|t| \geq 2$ (respectively, $w \in t\{t, f(t)\}\{t, f(t)\}^+$) is solvable in linear time (respectively, NP-complete) in the general case, is NP-complete for f non-erasing, and is solvable in $\mathcal{O}(n^2)$ time for f uniform.*

References

1. E. Czeizler, L. Kari, and S. Seki. On a special class of primitive words. *Theoretical Computer Science*, 411:617–630, 2010.
2. P. Gawrychowski, F. Manea, R. Mercas, D. Nowotka, and C. Tiseanu. Finding pseudo-repetitions. In N. Portier and T. Wilke, editors, *STACS*, volume 20 of *LIPIcs*, pages 257–268. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2013.
3. P. Gawrychowski, F. Manea, and D. Nowotka. Discovering hidden repetitions in words. In P. Bonizzoni, V. Brattka, and B. Löwe, editors, *CiE*, volume 7921 of *Lecture Notes in Computer Science*, pages 210–219. Springer, 2013.
4. J. Kärkkäinen, P. Sanders, and S. Burkhardt. Linear work suffix array construction. *J. ACM*, 53:918–936, 2006.
5. F. Manea, M. Müller, and D. Nowotka. The avoidability of cubes under permutations. In H.-C. Yen and O. H. Ibarra, editors, *Developments in Language Theory*, volume 7410 of *Lecture Notes in Computer Science*, pages 416–427. Springer, 2012.