Information Processing Letters 128 (2017) 54–57

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Two strings at Hamming distance 1 cannot be both quasiperiodic

Amihood Amir [a], Costas S. Iliopoulos [b], Jakub Radoszewski [b,*,1]

[a] Department of Computer Science, Bar-Ilan University, Ramat-Gan 52900, Israel
[b] Department of Informatics, King's College London, Strand, London WC2R 2LS, UK

ARTICLE INFO

Article history:
Received 26 September 2016
Received in revised form 25 August 2017
Accepted 28 August 2017
Communicated by P. Wong

Keywords:
Combinatorial problems
Combinatorics on words
Quasiperiodicity
Cover
Seed

ABSTRACT

We present a generalization to quasiperiodicity of a known fact from combinatorics on words related to periodicity. A string is called *periodic* if it has a period which is at most half of its length. A string w is called *quasiperiodic* if it has a non-trivial *cover*, that is, there exists a string c that is shorter than w and such that every position in w is inside one of the occurrences of c in w. It is a folklore fact that two strings that differ at exactly one position cannot be both periodic. Here we prove a more general fact that two strings that differ at exactly one position cannot be both quasiperiodic. Along the way we obtain new insights into combinatorics of quasiperiodicities.

1. Introduction

A *string* is a finite sequence of letters over an alphabet Σ. If w is a string, then by $|w| = n$ we denote its length, by $w[i]$ for $i \in \{1, \ldots, n\}$ we denote its i-th letter, and by $w[i..j]$ we denote a *factor* of w being a string composed of the letters $w[i] \ldots w[j]$ (if $i > j$, then it is the empty string). A factor $w[i..j]$ is called a *prefix* if $i = 1$ and a *suffix* if $j = n$.

An integer p is called a *period* of w if $w[i] = w[i + p]$ for all $i = 1, \ldots, n - p$. A string u is called a *border* of w if it is both a prefix and a suffix of w. It is a fundamental fact of string periodicity that a string w has a period p if and only if it has a border of length $n - p$; see [5,9]. If p is a period of w, $w[1..p]$ is called a *string period* of w. If w has a period p such that $p \le \frac{n}{2}$, then w is called *periodic*. In this case w has a border of length at least $\lceil \frac{n}{2} \rceil$.

For two strings w and w' of the same length n, we write $w =_j w'$ if $w[i] = w'[i]$ for all $i \in \{1, \ldots, n\} \setminus \{j\}$ and $w[j] \ne w'[j]$. This means that w and w' are at Hamming distance 1, where the Hamming distance counts the number of different positions of two equal-length strings. The following fact states a folklore property of string periodicity that we generalize in this work into string quasiperiodicity.

Fact 1. *Let w and w' be two strings of length n and $j \in \{1, \ldots, n\}$ be an index. If $w =_j w'$, then at most one of the strings w, w' is periodic.*

Fact 1 is, in particular, a consequence of a variant of Fine and Wilf's periodicity lemma that was proved by Berstel and Boasson in [2] in the context of partial words with one hole (a hole is a don't care symbol). For completeness we provide its proof in Section 4 without using the terms of partial words.

* Corresponding author.
E-mail addresses: amir@cs.biu.ac.il (A. Amir), costas.iliopoulos@kcl.ac.uk (C.S. Iliopoulos), jrad@mimuw.edu.pl (J. Radoszewski).
[1] The author is a Newton International Fellow.

http://dx.doi.org/10.1016/j.ipl.2017.08.005

a a b a a b a a a a b a a a b a a

Fig. 1. aabaa is a cover of aabaabaaaabaaabaa.

a b a a b a a a a b a a a b a a a

Fig. 2. aabaa is a seed of abaabaaaabaaabaaa.

We say that a string c *covers* a string w ($|w| = n$) if for every position $k \in \{1, \dots, n\}$ there exists a factor $w[i..j] = c$ such that $i \le k \le j$. Then c is called a *cover* of w; see Fig. 1. A string w is called *quasiperiodic* if it has a cover of length smaller than n.

A significant amount of work has been devoted to the computation of covers in a string. A linear-time algorithm finding the shortest cover of a string was proposed by Apostolico et al. [1]. Later a linear-time algorithm computing all the covers of a string was proposed by Moore and Smyth [10]. Breslauer [3] gave an on-line $O(n)$-time algorithm computing the cover array of a string of length n, that is, an array specifying the lengths of shortest covers of all the prefixes of the string. Li and Smyth [8] provided a linear-time algorithm for computing the array of longest covers of all the prefixes of a string that can be used to populate all the covers of every prefix. All these papers employ various combinatorial properties of covers.

Our main contribution is stated as the following theorem. As we have already mentioned, a periodic string has a border long enough to be the string's cover. Hence, a periodic string is also quasiperiodic, and Theorem 2 generalizes Fact 1.

Theorem 2. *Let w and w' be two strings of length n and $j \in \{1, \dots, n\}$ be an index. If $w =_j w'$, then at most one of the strings w, w' is quasiperiodic.*

The proof of Theorem 2 is divided into three sections. In Section 2 we restate several simple preliminary observations. Then, Section 3 contains a proof of a crucial auxiliary lemma which shows a combinatorial property of seeds that we use extensively in the main result. Finally, Section 4 contains the main proof.

2. Preliminaries

We say that a string s is a *seed* of a string w if $|s| \le |w|$ and w is a factor of some string u covered by s; see Fig. 2. Furthermore, s is called a *left seed* of w if s is both a prefix and a seed of w. Thus a cover of w is always a left seed of w, and a left seed of w is a seed of w. The notion of seed was introduced in [6] and efficient computation of seeds was further considered in [4,7].

In the proof of our main result we use the following easy observations that are immediate consequences of the definitions of cover and seed.

Observation 3. *Consider strings w and c.*

(a) *If c is a cover of w and $|c| \ge |w|/2$, then w is periodic with a period $|w| - |c|$.*
(b) *If c is a cover of w, then any cover of c is also a cover of w.*
(c) *If c is a seed of w, then c is a seed of every factor of w of length at least $|c|$.*
(d) *If w has a period p and a prefix of length at least p that has a cover c, then c is a left seed of w.*

A string w' is called a *cyclic shift* of a string w, both of length n, if there is a position $i \in \{1, \dots, n\}$ such that $w' = w[i+1..n]w[1..i]$. We denote this relation as $w \approx w'$. The following obviously holds.

Observation 4. *If w' is a cyclic shift of w, then w is a seed of w'.*

3. Auxiliary lemma

In the following lemma we observe a new property of the notion of seed. As we will see in Section 4, this lemma encapsulates the hardness of multiple cases in the proof of the main result.

Before we proceed to the lemma, however, let us introduce an additional notion lying in between periodicity and quasiperiodicity. We say that a string w of length n is *almost periodic* with period p if there exists an index $j \in \{1, \dots, n-p\}$ such that:

$$w[i] = w[i+p] \quad \text{for all } i = 1, \dots, n-p,\ i \ne j,$$
$$\text{and } w[j] \ne w[j+p].$$

In this case we refer to j as the *mismatch position*. Furthermore, if $w[1..b] =_j w[n-b+1..n]$ for an integer b, we say that each of these factors is an *almost border* of w of length b (and again refer to j as the mismatch position). We immediately observe the following.

Observation 5. *A string w of length n is almost periodic with period p and mismatch position j if and only if w has an almost border of length $n - p$ with mismatch position j.*

Example 1. The following string of length 19:

$\texttt{abaab ab}\underline{\texttt{a}}\texttt{ab ab}\underline{\texttt{b}}\texttt{ab abba}$

is almost periodic with period $p = 5$ and mismatch position $j = 8$ (the letters at positions j and $j + p$ are underlined). Hence, it has an almost border of length 14:

$\texttt{abaab ab}\underline{\texttt{a}}\texttt{ab abba} =_8 \texttt{abaab ab}\underline{\texttt{b}}\texttt{ab abba}.$

Lemma 6. *Let w and w' be two strings of length n and $j \in \{1, \dots, n\}$ be an index. If $w =_j w'$, then w is not a seed of w'.*

Proof. Assume to the contrary that w is a seed of w'. Let u be a string covered by w that has w' as a factor. Obviously, it suffices to consider two occurrences of w in u to cover all positions of the factor w': the leftmost one that covers $w'[n]$ and the rightmost one that covers $w'[1]$. Let α be the length of the longest suffix of w' that is a prefix of w, and let β be the length of the longest prefix of w' that is a suffix of w (these are the so-called longest overlaps between w' and w, and between w and w'). Thus we have $\alpha, \beta > 0$ and $\alpha + \beta \ge n$; see Fig. 3. From now on we