

Two strings at Hamming distance 1 cannot be both quasiperiodic

Amihod Amir¹, Costas S. Iliopoulos², and Jakub Radoszewski^{*,2,3}

¹ Department of Computer Science, Bar-Ilan University, Ramat-Gan 52900, Israel,
amir@cs.biu.ac.il

² Department of Informatics, King's College London, London, UK,
costas.ilopoulos@kcl.ac.uk

³ Institute of Informatics, University of Warsaw, Warsaw, Poland, jrad@mimuw.edu.pl

Abstract

We present a generalization of a known fact from combinatorics on words related to periodicity into quasiperiodicity. 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 $i \in \{1, \dots, n\}$ we denote its i -th letter, and by $w[i..j]$ we denote a *factor* of w being a string composed of the letters $w[i] \dots w[j]$ (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 $i = 1, \dots, n - p$. 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 [4, 8]. 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 $p \leq \frac{n}{2}$, 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 $i \in \{1, \dots, n\} \setminus \{j\}$ and $w[j] \neq w'[j]$. 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. For completeness we provide its proof in Section 4.

Fact 1. *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 periodic.*

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 $i \leq k \leq j$. 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 [9]. Breslauer [2] gave an

*The author is a Newton International Fellow.

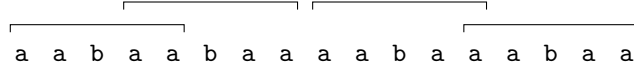


Figure 1: `aaba` is a cover of `aabaabaaaabaa`

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 [7] provided a linear-time algorithm for computing the array of longest covers of all the prefixes of a string. All these papers employ various combinatorial properties of covers.

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

Theorem 1. *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 1 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 $|s| \leq |w|$ 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 [5] and efficient computation of seeds was further considered in [3, 6].

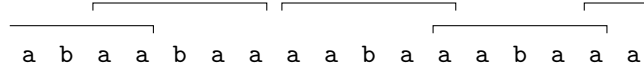


Figure 2: `aaba` is a seed of `abaabaaaabaa`

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 1. *Consider strings w and c .*

- (a) *If c is a cover of w and $|c| \geq |w|/2$, 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 2. *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:

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

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 3. *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:

abaab abaab abbab 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:

$$\text{abaab abaab abba} =_8 \text{abaab abbab abba}.$$

Lemma 1. *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 $\alpha + \beta \geq n$; see Fig. 3. From now on we assume that $\alpha \geq n/2$. The other case (i.e., $\beta \geq n/2$) is symmetric by reversing the strings w and w' . Let us denote $p = n - \alpha$.

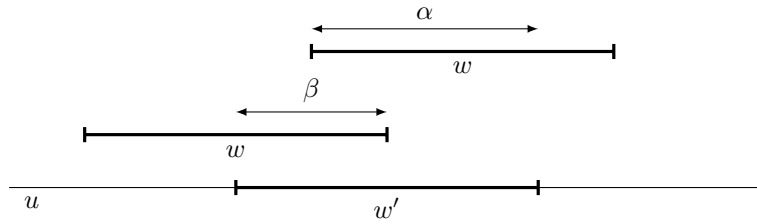


Figure 3: w is a seed of w' ; α and β are the longest overlaps between the two strings.

First consider the case when j satisfies $p < j \leq \alpha$. Then we have:

$$w'[1..\alpha] =_j w[1..\alpha] \quad \text{and} \quad w[1..\alpha] = w'[p + 1..n]$$

by the definitions of w' and α , respectively. Consequently, $w'[1..\alpha] =_j w'[p + 1..n]$. This means that w' has an almost border of length α with mismatch position j . By Observation 3, w' is almost periodic with period p and the same mismatch position.

The latter can be written equivalently as follows:

$$w'[1..p] \approx w'[2..p+1] \approx \dots \approx w'[j..j+p-1] \not\approx w'[j+1..j+p] \approx \dots \approx w'[\alpha+1..n]. \quad (1)$$

Recall that $w =_j w'$. This means that the same cyclic-shift relations hold for all corresponding factors of w that do not contain the symbol $w[j]$. Moreover, $w[1..\alpha] = w'[p+1..n]$, so $w[j] = w'[j+p] = w[j+p]$ (and $w[j] \neq w[j-p]$). This concludes that:

$$w[1..p] \approx w[2..p+1] \approx \dots \approx w[j-p..j-1] \not\approx w[j-p+1..j] \approx \dots \approx w[\alpha+1..n]. \quad (2)$$

Moreover, the inequalities satisfied by j in this case imply that $w'[1..p] = w[1..p]$ and $w'[\alpha+1..n] = w[\alpha+1..n]$. Hence, the conditions (1) and (2) conclude that there is no suffix of w of length at least p that would be a prefix of w' . Consequently, $\beta < p$, a contradiction.

We are left with two cases:

(A) $j \leq p$. In this case w has a border of length α :

$$w[1..\alpha] = w'[p+1..n] = w[p+1..n].$$

Consequently, w is periodic with period p . On the other hand, w' does not have the period p , since $w'[j+p] = w[j+p] = w[j] \neq w'[j]$. Moreover, $w'[1..p] \not\approx w[1..p] \approx w[i..i+p-1]$ for all i . In conclusion, there cannot exist a suffix of w of length at least p that would be a prefix of w' , i.e. $\beta < p$, a contradiction.

(B) $j > \alpha$. In this case w' has a border of length α :

$$w'[1..\alpha] = w[1..\alpha] = w'[p+1..n].$$

Consequently, w' is periodic with period p . On the other hand, w does not have the period p , since $w[j-p] = w'[j-p] = w'[j] \neq w[j]$. Moreover,

$$w[\alpha+1..n] \not\approx w'[\alpha+1..n] \approx w'[i..i+p-1]$$

for all i . In conclusion, there cannot exist a suffix of w of length at least p that would be a prefix of w' , i.e. $\beta < p$, a contradiction. \square

The following example illustrates the main case of the proof of the above lemma.

Example 2. Consider the following two strings of length 19:

$$w = \text{abaab ab}\underline{\text{b}}\text{ab abba abba}, \quad w' = \text{abaab ab}\underline{\underline{\text{a}}}\text{ab abba abba}.$$

We have $w =_8 w'$. The longest suffix of w' that is a prefix of w has length 14 (abaab abba abba). Hence, w' is almost periodic with period $19 - 14 = 5$ and mismatch position 8. Moreover, w is almost periodic with the same period and mismatch position $8 - 5 = 3$. We see that no prefix of w' of length at least 5 can be a suffix of w .

We use Lemma 1 as our key tool throughout the proof of the main result. As a consequence of Lemma 1 we obtain the following lemma that will be also useful in the main proof.

Lemma 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 there does not exist a string c that would be both a cover of w and a seed of w' .*

Proof. Consider an occurrence of c in w , $w[i..i+|c|-1]$, that covers the position j . Due to Observation 1c, the string $c' = w'[i..i+|c|-1]$ has c as a seed. We have $c' =_{j-i+1} c$, which contradicts Lemma 1. \square

4 Main Result

In this section we first present a proof of the folklore property of string periodicity (Fact 1) for completeness, and then proceed to the proof of our main result being a generalization of that fact (Theorem 1).

Proof (of Fact 1). Assume to the contrary that $w =_j w'$ and both strings are periodic. Let p and p' ($p, p' \leq \frac{n}{2}$) be the shortest periods of w and w' . Assume w.l.o.g. that $p \leq p'$. It suffices to prove the lemma in the case that w' is a square of length $2p'$ and $j \leq p'$. Let us define $w_1 = w[1..p']$ and $w_2 = w[p' + 1..2p']$. By the periodicity of w' , we see that $w_1 =_j w_2$.

We may assume that $j \leq \frac{p'}{2}$, as otherwise we may reverse both strings w_1, w_2 . Both w_1 and w_2 have period p , as they are factors of w (or the reversal of w), and their string periods of length p , further denoted by s_1 and s_2 , are cyclic shifts. Now consider any k such that $kp + 1 \leq j \leq (k + 1)p$ (it exists by the upper bound on the value of j). Then $w_1[kp + 1..(k + 1)p] = s_1$ and $w_2[kp + 1..(k + 1)p] = s_2$. This concludes that s_1 and s_2 differ at exactly one position $l = j - kp$, i.e., $s_2 =_l s_1$. However, s_1 is a cyclic shift of s_2 , hence a seed of s_2 by Observation 2. This contradicts Lemma 1. \square

Proof (of Theorem 1). Assume to the contrary that $w =_j w'$ and both strings are quasiperiodic. Let c and c' be the shortest covers of w and w' . W.l.o.g. we can assume that $|c| \leq |c'|$. We consider a few cases depending on the lengths of the covers:

- (A) $\frac{n}{2} \leq |c|$. By Observation 1a, the strings w, w' are both periodic. This contradicts Fact 1.
- (B) $|c| < \frac{n}{2} \leq |c'|$. Again by Observation 1a, w' is periodic with the period $p' = n - |c'| \leq \frac{n}{2}$. Assume w.l.o.g. that $j > \frac{n}{2}$. This means that the first half of w and w' , $w[1..[n/2]]$, has period p' and c is its left seed. There are three subcases:
 - (B1) $p' \leq |c|$. By Observation 1d, c is a left seed of w' . Therefore, w and w' contradict Lemma 2.
 - (B2) $|c| < p'$ and $p' < j \leq p' + |c|$. In this case the strings $s = w[p' + 1..p' + |c|]$ and $c = w'[p' + 1..p' + |c|] = w'[1..|c|]$ differ only at position $j - p'$. By Observation 1c applied to w , s has a seed c . This contradicts Lemma 1.
 - (B3) $|c| < p'$ and $j > p' + |c|$. Then c is a cover of $w'[1..p' + |c|]$, as it is a left seed of $w[1..p' + |c|] = w'[1..p' + |c|]$ and a suffix due to the period p' . Hence, by Observation 1d, c is a left seed of w' . Therefore, w and w' contradict Lemma 2.

From now on we assume that $|c|, |c'| < \frac{n}{2}$.

- (C) $c = c'$. This immediately contradicts Lemma 2.
- (D) $|c| = |c'|$ but $c \neq c'$. Let $m = |c|$; $m \leq n/2$. Then c is a border of w , and c' is a border of w' . As $c \neq c'$, it is not possible to change a single position in w such that both its prefix and its suffix of length m become c' .
- (E) $|c| < |c'|$. We consider three final subcases.
 - (E1) $|c| < j < n - |c| + 1$. This means that c is a border of w' , consequently a border of c' . However, c is not a cover of c' . Otherwise, by Observation 1b, c would be a cover of w' shorter than c' . Consider the factors $w[1..|c'|]$ and $w[n - |c'| + 1..n]$; note that they cover disjoint sets of positions. If $j \leq |c'|$, then $w[n - |c'| + 1..n] = w'[n - |c'| + 1..n] = c'$. The string c is a border of $c' = w[n - |c'| + 1..n]$ and a cover of w . Hence, by Observation 1c, c is a cover of c' . This contradicts the opposite observation that we have just made. Otherwise (if $j > |c'|$) we see that similarly c is a cover of $w[1..|c'|] = c'$, again a contradiction.
 - (E2) $j \geq n - |c| + 1$. This case is symmetric to the following case (D3) by reversing the strings w and w' .

(E3) $j \leq |c|$. As c' is a prefix of w' , this means that the prefix of c' of length $|c|$ is a string c_1 such that $c_1 =_j c$.

Note that $w[n - |c'| + 1..n] = w'[n - |c'| + 1..n] = c'$. The string c is a cover of w , therefore, by Observation 1c, c is a seed of the prefix c_1 of $w[n - |c'| + 1..n]$. This, however, contradicts Lemma 1.

The above cases include all the possibilities. This concludes the proof. \square

5 Conclusions

In this note we have proved that every two distinct quasiperiodic strings of the same length differ at more than one position. This bound is tight, as, for instance, for every even $n \geq 2$ the strings $\mathbf{a}^{n/2-1}\mathbf{b}\mathbf{a}^{n/2-1}\mathbf{b}$ and \mathbf{a}^n are both quasiperiodic and differ at exactly two positions.

Acknowledgements The authors thank Maxime Crochemore and Solon P. Pissis for helpful discussions.

References

- [1] Alberto Apostolico, Martin Farach, and Costas S. Iliopoulos. Optimal superprimitivity testing for strings. *Inf. Process. Lett.*, 39(1):17–20, 1991.
- [2] Dany Breslauer. An on-line string superprimitivity test. *Inf. Process. Lett.*, 44(6):345–347, 1992.
- [3] Michalis Christou, Maxime Crochemore, Costas S. Iliopoulos, Marcin Kubica, Solon P. Pissis, Jakub Radoszewski, Wojciech Rytter, Bartosz Szreder, and Tomasz Waleń. Efficient seeds computation revisited. In Raffaele Giancarlo and Giovanni Manzini, editors, *Combinatorial Pattern Matching - 22nd Annual Symposium, CPM 2011. Proceedings*, volume 6661 of *Lecture Notes in Computer Science*, pages 350–363. Springer, 2011.
- [4] Maxime Crochemore, Christophe Hancart, and Thierry Lecroq. *Algorithms on Strings*. Cambridge University Press, New York, NY, USA, 2007.
- [5] Costas S. Iliopoulos, D. W. G. Moore, and Kunsoo Park. Covering a string. *Algorithmica*, 16(3):288–297, 1996.
- [6] Tomasz Kociumaka, Marcin Kubica, Jakub Radoszewski, Wojciech Rytter, and Tomasz Waleń. A linear time algorithm for seeds computation. In Yuval Rabani, editor, *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012*, pages 1095–1112. SIAM, 2012.
- [7] Yin Li and William F. Smyth. Computing the cover array in linear time. *Algorithmica*, 32(1):95–106, 2002.
- [8] M. Lothaire. *Combinatorics on Words*. Addison-Wesley, Reading, MA., U.S.A., 1983.
- [9] Dennis Moore and William F. Smyth. Computing the covers of a string in linear time. In *SODA*, pages 511–515, 1994.