Information and Computation 209 (2011) 717–730

Contents lists available at ScienceDirect

Information and Computation

journal homepage: www.elsevier.com/locate/ic

An extension of the Lyndon–Schützenberger result to pseudoperiodic words ☆,☆☆

Elena Czeizler [1], Eugen Czeizler [2], Lila Kari, Shinnosuke Seki *

Department of Computer Science, The University of Western Ontario, London, Ontario, Canada N6A 5B7

ARTICLE INFO

Article history:
Received 26 October 2009
Revised 2 November 2010
Available online 11 January 2011

Keywords:
Pseudoperiodic word
(extended) Lyndon–Schützenberger equation
Pseudo-primitive word
Antimorphic involution
Non-trivial overlap

ABSTRACT

One of the particularities of information encoded as DNA strands is that a string u contains basically the same information as its Watson–Crick complement, denoted here as $\theta(u)$. Thus, any expression consisting of repetitions of u and $\theta(u)$ can be considered in some sense periodic. In this paper, we give a generalization of Lyndon and Schützenberger's classical result about equations of the form $u^l = v^n w^m$, to cases where both sides involve repetitions of words as well as their complements. Our main results show that, for such extended equations, if $l \geqslant 5$, $n, m \geqslant 3$, then all three words involved can be expressed in terms of a common word t and its complement $\theta(t)$. Moreover, if $l \geqslant 5$, then $n = m = 3$ is an optimal bound. These results are established based on a complete characterization of all possible overlaps between two expressions that involve only some word u and its complement $\theta(u)$, which is also obtained in this paper.

1. Introduction

Periodicity and primitiveness of words are fundamental properties in combinatorics on words and formal language theory. Their wide-ranging applications include pattern-matching algorithms (see, e.g., [1,2]) and data-compression algorithms (see, e.g., [3]). Sometimes motivated by their applications, these classical notions have been modified or generalized in various ways. A representative example is the "weak periodicity" of [4] whereby a word is called *weakly periodic* if it consists of repetitions of words with the same Parikh vector. This type of period was also called *abelian period* in [5]. Carpi and de Luca extended the notion of periodic words into that of periodic-like words according to the extendability of factors of a word [6]. Czeizler et al. have proposed the notion of *pseudo-primitiveness* (and pseudoperiodicity) of words in [7], motivated by the properties of information encoded as DNA strands.

DNA stores genetic information primarily in its single-stranded form as an oriented chain made up of four kinds of nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). Thus, a single-stranded DNA can be viewed as a word over the four-letter alphabet {A, C, G, T}. Due to the Watson–Crick complementarity of DNA strands, whereby A is complementary

☆ A short version of this paper was present at the 13th International Conference on Developments in Language Theory (DLT09) as: E. Czeizler, E. Czeizler, L. Kari, S. Seki, An extension of the Lyndon–Schützenberger result to pseudoperiodic words, in: Proceedings of DLT 09, LNCS, vol. 5583, 2009, pp. 183–194.

☆☆ This research was supported by Natural Sciences and Engineering Research Council of Canada Discovery Grant R2824A01, and Canada Research Chair Award to L.K.

* Corresponding author. Fax: +1 519 661 3515.
E-mail addresses: elena.czeizler@helsinki.fi (E. Czeizler), eugen.czeizler@aalto.fi (E. Czeizler), lila@csd.uwo.ca (L. Kari), sseki@csd.uwo.ca (S. Seki).
[1] Present address: Computational Systems Biology Laboratory, Faculty of Medicine, University of Helsinki, Finland.
[2] Present address: Department of Information and Computer Science, Aalto University, Aalto FI-00076, Finland.

0890-5401/$ - see front matter
doi:10.1016/j.ic.2011.01.001

to T, and C is complementary to G, single-stranded DNA molecules interact with each other. Indeed, two Watson–Crick complementary DNA single strands with opposite orientation will bind to each other by weak hydrogen bonds between their individual bases and form the well-known DNA double helix structure. In the process of DNA replication, a DNA double strand is separated into its two constituent single strands, each of which serves as a template for the enzyme called DNA polymerase. Starting from one end of a DNA single strand, DNA polymerase has the ability to build up, one nucleotide at a time, a new DNA strand that is perfectly complementary to the template, resulting in two copies of the DNA double strand. Thus, two DNA strands which are Watson–Crick complementary to each other can be considered "equivalent" in terms of the information they encode.

The fact that one can consider a DNA strand and its Watson–Crick complement "equivalent" led to natural and theoretically interesting extensions of various notions in combinatorics of words and formal language theory such as pseudo-palindrome [8], pseudo-commutativity [9], as well as hairpin-free and bond-free languages (e.g., [10–12]). Watson–Crick complementarity has been modeled mathematically by an antimorphic involution θ, i.e., a function that is an antimorphism ($\theta(uv) = \theta(v)\theta(u)$ for any words u, v), and an involution (θ^2 is the identity function). The aforementioned new concepts and notions are based on extending the notion of identity between words to that of "equivalence" between words u and $\theta(u)$, in the sense that an occurrence of $\theta(u)$ will be treated as another occurrence of u, albeit disguised by the application of θ.

In [7], a word w is said to be θ-*primitive* if we cannot find any word x that is strictly shorter than w such that w can be written as a combination of x and $\theta(x)$. For instance, if θ is the Watson–Crick complementarity then ATCG is θ-primitive, whereas TCGA is not because $\mathtt{TCGA} = \mathtt{TC}\theta(\mathtt{TC})$. The periodicity theorem of Fine and Wilf – one of the fundamental results on periodicity of words, see, e.g., [13,14] – was also extended as follows "For given words u and v, how long does a common prefix of a word in $\{u, \theta(u)\}^+$ and a word in $\{v, \theta(v)\}^+$ have to be, in order to imply that $u, v \in \{t, \theta(t)\}^+$ for some word t?".

In this paper, we continue the theoretical study of θ-primitive words by extending another central periodicity result, due to Lyndon and Schützenberger [15]. The original result states that, if the concatenation of two periodic words v^n and w^m can be expressed in terms of a third period u, i.e., $u^\ell = v^n w^m$, for some $\ell, m, n \geqslant 2$, then all three words u, v, and w can be expressed in terms of a common word t, i.e., $u, v, w \in \{t\}^+$ (see also [16] and Chapter 5 from [14] for some of its shorter proofs and [17,18] for some other generalizations). Replacing identity of words by the weaker notion of "equivalence" between words u and $\theta(u)$, for a given antimorphic involution θ, we define an extended Lyndon and Schützenberger equation as

$$u_1 \cdots u_\ell = v_1 \cdots v_n w_1 \cdots w_m, \tag{1}$$

where $u_1, \ldots, u_\ell \in \{u, \theta(u)\}$, $v_1, \ldots, v_n \in \{v, \theta(v)\}$, and $w_1, \ldots, w_m \in \{w, \theta(w)\}$ with $\ell, n, m \geqslant 2$. For this extended Lyndon and Schützenberger equation we ask the following question: "What conditions on ℓ, n, m imply that all three words u, v, w can be written as a combination of a word and its image under θ, i.e., $u, v, w \in \{t, \theta(t)\}^+$ for some word t?"

This paper gives a partial answer to the question that whenever $\ell \geqslant 5$, $n, m \geqslant 3$, Eq. (1) implies $u, v, w \in \{t, \theta(t)\}^+$ for some word t (Theorem 27), and that once either n or m becomes 2, we can construct u, v, w which satisfy Eq. (1), but such a word t does not exist (Examples 1 and 2). Therefore, for any $\ell \geqslant 5$, $n = m = 3$ is an optimal bound. In the case when $\ell = 3$ or $\ell = 4$, the problem of finding optimal bounds remains open, though Examples 1 and 2 work even in these cases. Our proofs are not generalizations of the methods used in the classical case, since one of the main properties used therein, i.e., the fact that the conjugate of a primitive word is still primitive, does not hold for θ-primitiveness any more.

Prior to the proof of the positive result, we characterize all non-trivial overlaps between two expressions $\alpha(v, \theta(v))$, $\beta(v, \theta(v)) \in \{v, \theta(v)\}^+$ for a θ-primitive word v. Formally speaking, we show that the equality $\alpha(v, \theta(v)) \cdot x = y \cdot \beta(v, \theta(v))$ with x and y shorter than v is possible, and we provide all possible representations of the involved words v, x, y (Theorem 14). Note that this result is in contrast to the classical case (where the two expressions involve only a word v, but not its image under θ).

The paper is organized as follows. In Section 2, we fix our terminology and recall some known results. In Section 3, we provide the characterization of all possible overlaps of the form $\alpha(v, \theta(v)) \cdot x = y \cdot \beta(v, \theta(v))$ with $\alpha(v, \theta(v))$, $\beta(v, \theta(v)) \in \{v, \theta(v)\}^+$ and x, y shorter than v. Finally, in Section 4 we provide our extension of Lyndon and Schützenberger's result.

2. Preliminaries

Here we introduce notions and notation used in the following sections. For details of each, readers are referred to [13,14].

Let Σ be a finite alphabet. We denote by Σ^* the set of all finite words over Σ, by λ the empty word, and by Σ^+ the set of all nonempty finite words. The catenation of two words $u, v \in \Sigma^*$ is denoted by either uv or $u \cdot v$. The *length* of a word $w \in \Sigma^*$ is denoted by $|w|$. We say that u is a *factor* (a *prefix*, a *suffix*) of v if $v = t_1 u t_2$ (resp. $v = u t_2$, $v = t_1 u$) for some $t_1, t_2 \in \Sigma^*$. We denote by $\mathrm{Pref}(v)$ (resp. $\mathrm{Suff}(v)$) the set of all prefixes (resp. suffixes) of the word v. We say that two words u and v overlap if $ux = yv$ for some $x, y \in \Sigma^*$ with $|x| < |v|$. An integer $p \geqslant 1$ is a *period* of a word $w = a_1 \ldots a_n$, with $a_i \in \Sigma$ for all $1 \leqslant i \leqslant n$, if $a_i = a_{i+p}$ for all $1 \leqslant i \leqslant n - p$.

A word $w \in \Sigma^+$ is called *primitive* if it cannot be written as a power of another word; that is, if $w = u^n$, then $n = 1$ and $w = u$. For a word $w \in \Sigma^+$, the shortest $u \in \Sigma^+$ such that $w = u^n$ for some $n \geqslant 1$ is called the *primitive root* of the word