Information and Computation 239 (2014) 222–236

Contents lists available at ScienceDirect

Information and Computation

www.elsevier.com/locate/yinco

The pseudopalindromic completion of regular languages

Szilárd Zsolt Fazekas [a], Florin Manea [b], Robert Mercaş [b,*], Kayoko Shikishima-Tsuji [c]

[a] Akita University, Department of Computer Science and Engineering, 1-1 Tegata Gakuen-machi, Akita City, 010-8502, Japan
[b] Kiel University, Department of Computer Science, D-24098 Kiel, Germany
[c] Tenri University, 1050 Somanouchi, Tenri, 632-8510, Japan

ARTICLE INFO

Article history:
Received 8 August 2013
Received in revised form 1 August 2014
Available online 28 September 2014

Keywords:
Pseudopalindromes
Pseudopalindromic completion
Pseudopalindromic iterated completion
Regular languages
Algorithms
Decidability

ABSTRACT

Pseudopalindromes are words that are fixed points for some antimorphic involution. In this paper we discuss a newer word operation, that of pseudopalindromic completion, in which symbols are added to either side of the word such that the new obtained words are pseudopalindromes. This notion represents a particular type of hairpin completion, where the length of the hairpin is at most one. We give precise descriptions of regular languages that are closed under this operation and show that the regularity of the closure under the operation is decidable.

1. Introduction

Palindromes are sequences which read the same starting from either end. Whenever the first half is the same as the complement of the second half read from right to left the sequence is called a pseudopalindrome. Besides their importance in combinatorial studies of strings, mirrored complementary sequences occur frequently in DNA and are often found at functionally interesting locations such as replication origins or operator sites. Already in the 1950's it was recognised that pseudopalindromic regions of DNA can exist in a cruciform structure with intrastrand base pairing of the self-complementary sequence, i.e., if a pseudopalindromic sequence occurs in a double strand, then pulling apart the two strands at the middle of the pseudopalindrome one can perform a "transfer-twist" in which each strand twists about itself, reducing the energy needed to separate the strands.

A similar phenomenon is when a single strand of DNA curls back on itself to become self-complementary, after which a polymerase chain reaction can extend the "shorter" end to generate a complete double strand, the result being a partial double helix with a bend in it. The structure is called a hairpin or stem-loop, and it is an important building block of many RNA secondary structures. Several operations on sequences were introduced which are either directly motivated by the biological phenomenon called stem-loop completion, or are very similar in nature to it. The mathematical hairpin concept introduced by Păun et al. in [1] refers to a word in which some suffix is the mirrored complement of a middle factor of the word. The hairpin completion operation, which extends such a word into a pseudopalindrome with a non-matching part in the middle was thoroughly investigated in [2–9]. Most basic algorithmic questions about hairpin completion have been

* Corresponding author.
E-mail addresses: szilard.fazekas@gmail.com (S.Z. Fazekas), flm@informatik.uni-kiel.de (F. Manea), rgm@informatik.uni-kiel.de (R. Mercaş), tsuji@sta.tenri-u.ac.jp (K. Shikishima-Tsuji).

http://dx.doi.org/10.1016/j.ic.2014.09.001
0890-5401/

answered (see, e.g. Cheptea et al. [2] and Diekert et al. [3]) with a noteworthy exception: "given a word, can we decide whether the iterated application of the operation leads to a regular language?". There exist however particular cases of this concept, such as the bounded hairpin completion introduced by Ito et al. in [10], where also this latter problem was settled by Kopecki in [11].

Another operation related to our topic is the iterated palindromic closure which was first introduced by de Luca in the study of the Sturmian words in [12] and later generalised to pseudopalindromes by de Luca and De Luca in [13]. This operator allows one to construct words with infinitely many pseudopalindromic prefixes, called pseudostandard words. In this context the newly obtained words have length greater than the original ones, but are minimal among all pseudopalindromes that have the original word as prefix or suffix. On the same page, in [14] Mahalingam and Subramanian propose the study of some similar operation, that of pseudopalindromic completion of a word. The operation considers in this form all possible ways of extending the word into a pseudopalindrome, thus producing in one step of an application of the operation an infinite set from any starting word. It is worth noting that together with the notion of θ-inverse that refers to the shortest such completion and was introduced in the same paper, the latter concept represents a generalisation of the pseudopalindromic closure.

The operation studied here, is yet another type of pseudopalindromic completion. Although having the same name, it differs from the pseudopalindromic completion in [14] in that we require the word to have a pseudopalindromic prefix or suffix in order to be completed. In comparison with the (iterated) pseudopalindromic closure in [12] which considers the unique shortest word that completes the starting string into a pseudopalindrome, as we shall observe, here we take all possible extensions. However, this choice of restriction was not randomly picked, the subject of this work being closest in nature to the first presented operation. Since in the biological phenomenon serving as inspiration, the hairpin's length in the case of stable bindings is optimally limited (approximatively 4–8 base-pairs according to Slama-Schwok et al. in [15]) it is natural to consider completions with bounded middle part. As we will see our operation is in fact a rather restricted form of the hairpin completion (we do not allow for non-matching middles), and the questions asked are also a subset of problems considered for the initial operation.

After presenting the notions and results needed for our treatise, in Section 3 we state some simple one-step completion results. In Section 4 we gradually build the characterisation of regular languages which stay regular under the iterated application of pseudopalindromic completion. Section 5 is a collection of algorithmic results on this operation such as the membership problem for the iterated completion of a word and that of a language, the completion distance between two input words and decision methods regarding the preservation of regularity within iterated completion. We conclude our work with Section 6 where we present some further remarks regarding this operation.

2. Preliminaries

We assume the reader to be familiar with fundamental concepts from Formal Language Theory, such as the classes of the Chomsky hierarchy, finite automaton, regular expressions (e.g., see the textbook by Harrison [16]), as well as fundamental concepts from combinatorics on words (e.g., see the Lothaire textbook [17]).

Let Σ be a non-empty finite *alphabet* with *letters* as elements. A sequence of letters constitutes a *word* $w \in \Sigma^*$ and we denote by ε the *empty word*.

The *length* of a finite word w is the number of not necessarily distinct symbols it consists of and is denoted by $|w|$. The ith symbol we write as $w[i]$ and use the notation $w[i..j]$ to refer to the part of a word starting at the ith and ending at the jth position.

Words together with the operation of concatenation form a free monoid, which is usually denoted by Σ^* for an alphabet Σ. Any subset of Σ^* is called a language. By Σ^+ we denote the set of non-empty words over Σ.

Repeated concatenation of a word w with itself is denoted by w^i for integers $i \geq 0$ with i representing a *power*. Furthermore, w is said to be *primitive* if there exists no non-empty word u such that $w = u^j$ for some integer $j > 1$. Otherwise, we call w a *repetition* and the smallest such u its *root* (note that in this case the word u is primitive).

A word u is a *factor* of w if there exist integers i, j with $1 \leq i, j \leq |w|$ such that $u = w[i..j]$. We say that u is a *prefix* of w whenever we can fix $i = 1$ and denote this by $u \leq_p w$. If $j < |w|$, then the prefix is called *proper*. *Suffixes* are the corresponding concept reading from the back of the word to the front. A word w has a positive integer k as a *period* if for all i, j such that $i \equiv j (\mathrm{mod} k)$ we have $w[i] = w[j]$, whenever both $w[i]$ and $w[j]$ are defined.

A central concept to this work is *palindromicity* in the general sense. For a word $w \in \Sigma^*$ we denote by w^R its *reversal*, that is $w[|w|]w[|w|-1]\dots w[1]$. If $w = w^R$, the word is called a *palindrome*. Let $\mathcal{P}al(L)$ be the set of all palindromes of a language $L \subseteq \Sigma^*$ and $\mathcal{P}al_\Sigma$ be the language of all palindromes over Σ (when the alphabet is clear from the context we shall drop the Σ and denote this set by $\mathcal{P}al$).

We can generalise the previous definition by using an arbitrary antimorphic involution instead of the reversal. To this end, let θ be an *antimorphic involution*, i.e., $\theta : \Sigma^* \to \Sigma^*$ is a function such that $\theta(\theta(a)) = a$ for all $a \in \Sigma$ and $\theta(uv) = \theta(v)\theta(u)$ for all $u, v \in \Sigma^+$. Then, w is a *θ-pseudopalindrome* if $w = \theta(w)$. To make notation simpler, we write $\overline{u}$ for $\theta(u)$ whenever θ is understood from the context. As an example, for the antimorphism θ with $\theta(a) = b$ and $\theta(b) = a$, the word $aabb = \theta(aabb) = \theta(b)\theta(b)\theta(a)\theta(a)$ is a pseudopalindrome. The language of all θ-pseudopalindromes, when the alphabet Σ and θ are fixed, is denoted by $\mathcal{PP}al$.