Towards the Sequence Design Preventing Pseudoknot Formation

Lila Kari and Shinnosuke Seki

Department of Computer Science, University of Western Ontario,
London, Ontario, Canada, N6A 5B7
{lila,sseki}@csd.uwo.ca

Abstract. This paper addresses a pseudoknot-freeness problem of DNA and RNA sequences, motivated by biomolecular computing. Watson-Crick (WK) complementarity forces DNA strands to fold into themselves and form so-called secondary structures, which are usually undesirable for biomolecular computational purposes. This paper studies pseudoknot-bordered words, a mathematical formalization of a common secondary structure, the pseudoknot. We obtain several properties of WK-pseudoknot-bordered and -unbordered words. One of the main results of the paper is that a sufficient condition for a WK-pseudoknot-unbordered word u to result in all words in u^+ being WK-pseudoknot-unbordered is for u not to be a primitive word.

1 Introduction

Adleman's first biomolecular computing experiment [1] has shown that biochemical properties of DNA such as Watson-Crick (WK) complementarity make it possible to solve computational problems, such as NP-complete problems entirely by DNA manipulation in test tubes. In DNA computing, information is encoded into DNA by a coding scheme mapping the original alphabet onto DNA single strands over {Adenine (A), Guanine (G), Cytosine (C), Thymine (T)}. A computation consists of a succession of *bio-operations* [2] based on *base-pairing* and the others. A can chemically bind to T, while C can similarly bind to G. (Note that T is replaced by U in the case of ribonucleic acid (RNA), and that U is complementary to both C and G.) Bases that can thus bind are called *Watson/Crick (WK) complementary*, and two DNA single strands with opposite orientation and with WK complementary bases at each position can bind to each other to form a *DNA double strand*.

Watson-Crick complementarity often makes a single-stranded structure fold into a high-dimensional (partially double-stranded) structure that is optimal in terms of biochemical determinants like Gibbs free-energy [3]. *In vivo* the secondary structures of nucleic acids have a significant role in determining their biochemical functions. On the other hand, *in vitro* biomolecular computing often considers them as disadvantages because it is very likely that the secondary structure formation of a DNA/RNA strand will prevent it from interacting with other DNA/RNA strands in the expected, pre-programmed way. Thence, many

Y. Suzuki et al. (Eds.): IWNC 2007, PICT 1, pp. 101–110, 2009.

studies exist on how to free sequence sets from secondary structures [4], [5], [6], [7], [8], [9], [10].

From the intramolecular point of view, most of these studies investigate the question of how to design sets of DNA/RNA strands that are "free of" the hairpin structure, known as the most common secondary structure. A hairpin structure can be formally modelled by using the notion of an antimorphic involution. An involution is a function f such that f^2 equals the identity, and an antimorphism f over an alphabet Σ is a function such that $f(uv) = f(v)f(u)$ for all words $u, v \in \Sigma^*$. An antimorphic involution is thus the mathematical formalization of the notion of DNA single-strand Watson/Crick complementarity. Indeed, the WK complement of a single DNA strand is the reverse, complement of the original strand. Using this formalization, a hairpin can be described as $z\beta\theta(z)$ as indicated in Fig. 1 (b), and modelled by the notion of a θ-bordered word [9]. In other words, a set of θ-unbordered words is guaranteed to be hairpin-free, and as such, results obtained in [9] on θ-bordered words for antimorphic involution θ are of practical significance.

In this paper, we take a similar approach to modeling and structure-freeness problems that ensures that no pseudoknots structures will be formed. Pseudoknots are a generic term of a cross-dependent structure that are formed primarily by RNA strands. An example of the simplest and most typical pseudoknots is shown in Figure 1, (a). An example of a pseudoknot found in *E.Coli* tmRNA is depicted in Figure 2. This depicts a pseudoknot formed by a strand $u = \rho x\alpha y\gamma\theta(x)\delta\theta(y)\sigma$ where x and $\theta(x)$ respectively y and $\theta(y)$ bind to each other. In this paper we consider the simpler case wherein $\rho = \alpha = \delta = \sigma = \lambda$, *i.e.*, we investigate strands of the form $u = xy\gamma\theta(x)\theta(y)$ where θ is an antimorphic involution function.

We namely generalize the notion of θ-(un)bordered word to that of θ-*pseudoknot-(un)bordered word*. A word is called θ-pseudoknot-bordered if it has a prefix whose image under the composition of the cyclic permutation and θ is its suffix. Formally speaking, a word w is θ-pseudoknot-bordered if $w = xy\alpha = \beta\theta(yx)$ for some words x, y, α, and β. In the case where θ is an antimorphic involution, this indeed is a formal model for simple pseudoknots since $\theta(yx) = \theta(x)\theta(y)$ holds.

This paper is organized as follows: After basic definitions and notations in Sec. 2, we define the notion of θ-pseudoknot-bordered words in Sec. 3 and prove some basic properties. We also show that the notion of θ-pseudoknot-border generalizes the notion of a θ-border. Sec. 4 concludes this paper by providing a counterintuitive result, Corollary 4, which proves that the sufficient condition for a θ-pseudoknot-unbordered word u to satisfy the condition that all words in u^+ have the same property turns out to be that u be not primitive. Proofs of the results in this paper can be found in [11].

2 Preliminaries

This section introduces basic notions of formal language theory and algebra. For details of each notion contained in this section, we refer the reader to [13], [14].

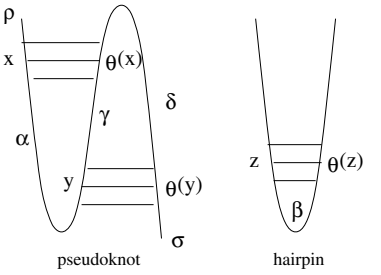

Fig. 1. Examples of a) a pseudoknot and b) a strand with two hairpins

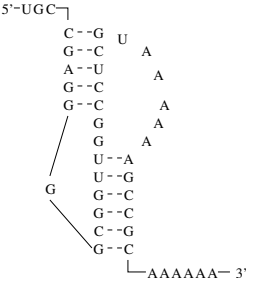

Fig. 2. A pseudoknot found in E. Coli tmRNA. Here $u = \rho x \alpha y \gamma \theta(x) \delta \theta(y) \sigma$, where $\rho =$ `UGC`, $x =$ `CGAGG`, $\alpha =$ `G`, $y =$ `GCGGUU`, $\gamma =$ `GG`, $\delta =$ `UAAAAA`, $\sigma =$ `AAAAAA`, and x and $\theta(x)$ respectively y and $\theta(y)$ bind to each other. From [12].

Let Σ^* (resp. Σ^+) be the free monoid (resp. free semigroup) generated by a finite alphabet Σ with the concatenation operation. The identity element of Σ^* is denoted by λ and as such, $\Sigma^+ = \Sigma^* \setminus \{\lambda\}$. An element of Σ^* is called a word. Hereafter words will be denoted by lower-case letters such as x, y, α, β. For a word $w \in \Sigma^*$, $|w|$ denotes the length of w. Let u and w be words over Σ. We say that u is a *prefix* of w if there exists $v \in \Sigma^*$ such that $w = uv$; Similarly, u is a *suffix* of w if $w = vu$ for some $v \in \Sigma^*$. Let $\mathrm{Pref}(w)$ and $\mathrm{Suff}(w)$ be the set of all prefixes and that of all suffixes of w, respectively.

A word $w \in \Sigma^+$ is *primitive* if $w = u^p$ with $u \in \Sigma^+$ implies $p = 1$. It is a well known fact [15], that for any word $w \in \Sigma^*$, there exist a unique primitive word, which is denoted by $\sqrt{w}$ and called the primitive root of w, and a unique positive integer k such that $w = (\sqrt{w})^k$.

Let θ be a mapping on a set S. If $a = \theta(a)$ for all $a \in S$, then θ is called the identity function or simply the identity. An *involution* is a mapping whose square is the identity. In this paper we consider two mappings on Σ^*: a d-morphism and a cyclic permutation. A *d-morphism* on Σ^* is a generic term used to refer to either a morphism or an antimorphism on Σ^*. A mapping θ from Σ^* to itself is defined as a *morphism* (resp. *antimorphism*) on Σ^* if and only if $\theta(xy) = \theta(x)\theta(y)$ (resp. $\theta(xy) = \theta(y)\theta(x)$) for all $x, y \in \Sigma^*$. For a d-morphic involution θ, a word $w \in \Sigma^*$ is called *θ-palindrome* if and only if $w = \theta(w)$. We denote by P_θ the set

of all θ-palindromes over Σ^*. For a word $w \in \Sigma^*$ such that $w = xy$ for $x, y \in \Sigma^*$, yx is called a *cyclic permutation* of w. The set of all cyclic permutations of w is denoted by $\mathrm{cp}(w)$; that is, $\mathrm{cp}(w) = \{yx \mid w = xy, x, y \in \Sigma^*\}$.

With applications to the function θ being the Watson-Crick complementarity of DNA sequences in mind, hereafter we shall deal only with non-identity mappings. Thus, this paper excludes singleton alphabet sets, on which there does not exist any non-identity mapping, that is, we assume $|\Sigma| \geq 2$.

3 θ-Pseudoknot-Bordered Words

In this section we introduce the notion of *θ-pseudoknot-bordered word*, for a morphic or antimorphic involution θ. This notion is a formalization of the biological concept of pseudoknot. Indeed, if θ is the Watson-Crick involution, then a θ-pseudoknot-bordered word represents a DNA/RNA strand that forms a pseudoknot as pictured in Fig. 1.

In addition, the notion of θ-pseudoknot-bordered word represents a proper generalization of the classical notion of a bordered word. A non-empty word u is called *bordered* [16] if there exists a non-empty word v that is both a prefix and a suffix of u. An unbordered word is a non-empty word that is not bordered.

The first step towards generalizing the notion of a bordered word was in [9], where the concept of θ-bordered word was first defined, that generalized the identity function by replacing it with a d-morphic involution θ. Here we propose the next step in this direction by employing a cyclic permutation to further extend the notion of a θ-bordered word to that of a *θ-pseudoknot-bordered word*.

Definition 1. *([9]) Let θ be a d-morphic involution, and $v, w \in \Sigma^*$. Then,*

1. *$v \leq_p w$ if and only if $w \in v\Sigma^*$. The word v is a prefix of the word w.*
2. *$v \leq_s^\theta w$ if and only if $w \in \Sigma^*\theta(v)$. The word v is a θ-suffix of the word w.*
3. *$\leq_d^\theta = \leq_p \cap \leq_s^\theta$. If $u, v \in \Sigma^*$ and $v \leq_d^\theta u$ we say that v is a θ-border of u. A word $w \in \Sigma^+$ is said to be θ-bordered if there exists $v \in \Sigma^+$ such that $v <_d^\theta w$, i.e., $w = v\alpha = \beta\theta(v)$ for some $\alpha, \beta \in \Sigma^+$. A non-empty word which is not θ-bordered is called θ-unbordered.*
4. *$v <_p w$ if and only if $w \in v\Sigma^+$. v is a proper prefix of w.*
5. *$v <_s^\theta w$ if and only if $w \in \Sigma^+\theta(v)$. v is a proper θ-suffix of w.*
6. *$<_d^\theta = <_p \cap <_s^\theta$. If $u, v \in \Sigma^*$ and $v <_d^\theta u$ we say that v is a proper θ-border of u.*
7. *For $w \in \Sigma^+$, $L_d^\theta(w) = \{v \mid v \in \Sigma^*, v <_d^\theta w\}$. $L_d^\theta(w)$ denotes the set of all proper θ-borders of a nonempty word w.*
8. *$D_\theta(i) = \{w \mid w \in \Sigma^+, |L_d^\theta(w)| = i\}$. $D_\theta(i)$ denotes the set of all nonempty words that have exactly i θ-borders.*

We now generalize this definition with the goal of defining the notion of the θ-pseudoknot-bordered word. This is accomplished by introducing a cyclic permutation.

Definition 2. *Let θ be a d-morphic involution, and $v, w \in \Sigma^*$. Then,*

1. $v \leq_{\mathrm{cs}}^{\theta} w$ *if and only if there exists* $v' \in \mathrm{cp}(v)$ *such that* $v' \leq_{\mathrm{s}}^{\theta} w$. *In other words,* $v \leq_{\mathrm{cs}}^{\theta} w$ *iff* $v = xy$, $x, y \in \Sigma^*$, *and* $w = \beta\theta(yx)$. v *is called a θ-pseudoknot-suffix of w.*
2. $\leq_{\mathrm{cd}}^{\theta} = \leq_{\mathrm{p}} \cap \leq_{\mathrm{cs}}^{\theta}$. v *is said to be a θ-pseudoknot-border of w if* $v \leq_{\mathrm{cd}}^{\theta} w$, *i.e., there exist* $x, y \in \Sigma^*$ *such that* $v = xy$ *and* $w = xy\alpha = \beta\theta(yx)$ *for some* $\alpha, \beta \in \Sigma^*$. *A non-empty word w is said to be θ-pseudoknot-bordered if w has a non-empty θ-pseudoknot-border. A non-empty word which is not θ-pseudoknot-bordered is called θ-pseudoknot-unbordered.*
3. $\mathrm{L}_{\mathrm{cd}}^{\theta}(w) = \{v \mid v \in \Sigma^*, v \leq_{\mathrm{cd}}^{\theta} w\}$. $\mathrm{L}_{\mathrm{cd}}^{\theta}(w)$ *denotes the set of all θ-pseudoknot-borders of a nonempty word w.*
4. $\mathrm{K}_{\theta}(i) = \{w \mid |w \in \Sigma^+, \mathrm{L}_{\mathrm{cd}}^{\theta}(w)| = i\}$. $\mathrm{K}_{\theta}(i)$ *denotes the set of all nonempty words that have exactly i θ-pseudoknot-borders.*

As in the case of θ-bordered words, the empty word λ is a θ-pseudoknot-border of any word, *i.e.*, $\forall w \in \Sigma^*, \lambda \in \mathrm{L}_{\mathrm{cd}}^{\theta}(w)$. Indeed, Definition 2 (2) allows the cases $v = xy = \lambda$ and $w = \alpha = \beta$. Thus, a word in $\mathrm{K}_{\theta}(1)$ has no θ-pseudoknot-borders other than λ. $\mathrm{K}_{\theta}(1)$ is the set of all θ-pseudoknot-unbordered words.

Note also that it is possible that a word w has itself as its θ-pseudoknot-border, as shown by the following example.

Example 1. Let θ be an antimorphic involution on Σ^*, and let $a, b \in \Sigma$ such that $\theta(a) = b$ and $\theta(b) = a$. Consider $u = ababbbaa$, which can be factorized into two θ-palindromes $abab$ and $bbaa$ and thus u is one of its θ-pseudoknot-borders. It is easy to see that the only other θ-pseudoknot-border is λ, and thus $u \in \mathrm{K}_{\theta}(2)$. □

Lastly, note that the definitions of $\mathrm{L}_{\mathrm{d}}^{\theta}(w)$ and $\mathrm{L}_{\mathrm{cd}}^{\theta}(w)$ are different in that the former equals the set of all the proper θ-borders while the latter can include also w, if w is a θ-pseudoknot-border of itself. This scenario is different from the classical case of θ as well as the permutation used being the indentity, wherein all words are automatically borders of themselves. We found our proposed definition to be more natural in the case of θ-pseudoknot-borders, since only some words w are θ-pseudoknot-borders of themselves while others are not. This implies however that, while all other proposed notions are strict generalizations of both the θ-border notions and the classical border notion, $\mathrm{L}_{\mathrm{cd}}^{\theta}(w)$ does not strictly generalize $\mathrm{L}_{\mathrm{d}}^{\theta}(w)$ and $\mathrm{L}_{\mathrm{d}}(w)$. This was a deliberate choice of definition on our part since a) this definition is more natural and b) all results that we obtained in this paper hold for the other definition choice as well, either unchanged or augmented with a weak additional condition. For example, Proposition 2 holds even if we define $\mathrm{L}_{\mathrm{cd}}^{\theta}$ to be the set of all proper θ-pseudoknot-bordered words, if we require, in addition, that u cannot be factorized into two θ-palindromes, *i.e.*, there exist no $x, y \in \mathrm{P}_{\theta}$ such that $u = xy$.

In the sequel, we will employ the expression "*xy is a θ-pseudoknot-border of w*" to mean "v is a θ-pseudoknot-border of w such that $v = xy$ and $w = xy\alpha = \beta\theta(yx)$ for some $x, y, \alpha, \beta \in \Sigma^*$".

To begin with, we provide some immediate consequences of Definition 2.

Corollary 1. *If θ is a d-morphic involution on Σ^*, the followings hold.*

1. *If a word has some θ-pseudoknot-border of length n, then for every $a \in \Sigma$, the number of letters a in its prefix of length n must be equal to the number of letters $\theta(a)$ in its suffix of length n.*
2. *For all $a \in \Sigma$ such that $a \neq \theta(a)$, $a^+ \subseteq \mathrm{K}_\theta(1)$.*
3. *If $xy \leq^\theta_{\mathrm{cd}} w^n$ and $|w^{m-1}| < |xy| \leq |w^m|$, then $xy \leq^\theta_{\mathrm{cd}} w^k$ for all k with $m \leq k \leq n$.*

Example 2. Let $\Sigma = \{a, b\}$, $w = aababbb$, and θ be the antimorphic involution such that $\theta(a) = b$ and $\theta(b) = a$. Then $L^\theta_{cd}(w) = \{\lambda, a, aa, aaba\}$. In particular, setting $x = aab$ and $y = a$ confirms that $aaba \leq^\theta_{\mathrm{cd}} w$. □

Recall that a language L is said to be *dense* if $\forall w \in \Sigma^*, L \cap \Sigma^* w \Sigma^* \neq \emptyset$.

Lemma 1. *Let θ be a d-morphic involution on Σ^*. Then $\mathrm{K}_\theta(1)$, the set of all θ-pseudoknot-unbordered words over Σ^*, is a dense set.*

The following lemma and proposition show that if a word is θ-bordered, then it is θ-pseudoknot-bordered.

Lemma 2. *Let θ be a d-morphic involution on Σ^* and $w \in \Sigma^*$. Then $\mathrm{L}^\theta_{\mathrm{d}}(w) \subseteq \mathrm{L}^\theta_{\mathrm{cd}}(w)$ holds.*

Proof. Let $v \in \mathrm{L}^\theta_{\mathrm{d}}(w)$. If $v = \lambda$, then by definition, $v \in \mathrm{L}^\theta_{\mathrm{cd}}(w)$; otherwise $w = v\alpha = \beta\theta(v)$ for some $\alpha, \beta \in \Sigma^*$. This means that we can split v into $x = v$ and $y = \lambda$ so as to satisfy the condition of v being a θ-pseudoknot-border of w, *i.e.*, $w = v\lambda\alpha = \beta\theta(\lambda v)$. This implies that $v \in \mathrm{L}^\theta_{\mathrm{cd}}(w)$.

Note that there exists a word $w \in \Sigma^*$ and a *d*-morphic involution θ for which $\mathrm{L}^\theta_{\mathrm{d}}(w)$ is strictly included in $\mathrm{L}^\theta_{\mathrm{cd}}(w)$.

Example 3. Let $\Sigma = \{a, b\}$, $w = aababbb$, and θ be a *d*-morphic involution satisfying $\theta(a) = b$ and $\theta(b) = a$. Whether θ is morphic or antimorphic involution, $\mathrm{L}^\theta_{\mathrm{d}}(w) = \{\lambda, a, aa\}$ but $\mathrm{L}^\theta_{\mathrm{cd}}(w) = \{\lambda, a, aa, aaba\}$. □

Although, in this example, $\mathrm{L}^\theta_{\mathrm{cd}}(w)$ for a morphic involution θ and $\mathrm{L}^\theta_{\mathrm{cd}}(w)$ for an antimorphic involution θ are the same, that is not always the case as indicated in the following examples:

Example 4. Let us consider Σ and θ as in Example 3, a word $w = aabbabaababb$, and its prefix $w_p = aabbab$. When θ is morphic, we can decompose w_p into $x = aa$ and $y = bbab$ such that $\theta(yx)$ becomes the suffix of w. Thus, $aabbab \in \mathrm{L}^\theta_{\mathrm{cd}}(w)$ for the morphism θ. On the other hand, $aabbab \notin \mathrm{L}^\theta_{\mathrm{cd}}(w)$ for an antimorphic θ. □

Example 5. Let us consider Σ and θ as in Example 3, a word $w' = aabbabbbabaa$, and its prefix $w'_p = aabbab$. When θ is antimorphic, we can decompose w'_p into $x = aa$ and $y = bbab$ such that $\theta(yx)$ becomes the suffix of w'. Therefore, $aabbab \in \mathrm{L}^\theta_{\mathrm{cd}}(w')$ for the antimorphism θ. On the other hand, $aabbab \notin \mathrm{L}^\theta_{\mathrm{cd}}(w')$ for a morphic θ. □

Proposition 1. *Let θ be a d-morphic involution on Σ^*. Then* $\mathrm{K}_\theta(1) \subseteq \mathrm{D}_\theta(1)$.

For an antimorphic involution θ, note that the inclusion relation of Proposition 1 holds properly, i.e. $\mathrm{K}_\theta(1) \subsetneq \mathrm{D}_\theta(1)$, as shown in the following example.

Example 6. Let $w = aba$ and θ be an antimorphic involution mapping a to b and vice versa. Suppose $w \notin \mathrm{D}_\theta(1)$. Then w should be of the form $a\Sigma^*\theta(a)$ [9]. However, w does not end with $\theta(a)$, and we conclude that $w \in \mathrm{D}_\theta(1)$. On the other hand, $w \notin \mathrm{K}_\theta(1)$ because $w = xya = a\theta(yx)$ for $x = a$ and $y = b$. □

4 Primitive and θ-Pseudoknot-Unbordered Words

This section addresses the question of whether or not the Kleene closure of a θ-pseudoknot-unbordered word contains only θ-pseudoknot-unbordered words. In other words, if $u \in \mathrm{K}_\theta(1)$ we are asking whether or not the inclusion $u^+ \subseteq \mathrm{K}_\theta(1)$ holds. This question was solved positively for θ-unbordered words in [9], that is, if $u \in D_\theta(1)$, then $u^+ \subseteq D_\theta(1)$. In contrast, in this section we answer in the negative the question for the case of θ-pseudoknot-unbordered words. Moreover, we provide a sufficient condition for a θ-pseudoknot-unbordered word $u \in \mathrm{K}_\theta(1)$ to satisfy $u^+ \subseteq \mathrm{K}_\theta(1)$. Unexpectedly, the condition is that u is not primitive (Corollary 4).

To begin with, we provide a necessary and sufficient condition for a word to be θ-pseudoknot-unbordered.

Lemma 3. *Let θ be an antimorphic involution on Σ^*. Then for $u \in \Sigma^+$, u is θ-pseudoknot-unbordered if and only if* $\theta(\mathrm{cp}(\mathrm{Pref}(u))) \cap \mathrm{Suff}(u) = \emptyset$.

The following lemma will be used as a tool to prove that a nonempty word $w \in \Sigma^+$ is θ-pseudoknot-bordered by *reductio ad absurdum*.

Lemma 4. *Let θ be an antimorphic involution, and x and y be θ-palindromes. If a word $u \in \Sigma^+$ has xy as both its prefix and suffix, then u is θ-pseudoknot-bordered, i.e., $u \notin \mathrm{K}_\theta(1)$.*

Next, we relate the property of a word w being θ-pseudoknot-unbordered to the fact that u^k is θ-pseudoknot-bordered for some integer $k > 1$. This result relates to the following results obtained for the particular case of the θ-bordered words in [9].

Lemma 5. *([9]) Let θ be an antimorphism on Σ^* and let $u \in \Sigma^+$. Then* $\theta(\mathrm{Pref}(u)) \cap \mathrm{Suff}(u) = \emptyset$ *if and only if* $\theta(\mathrm{Pref}(u^+)) \cap \mathrm{Suff}(u^+) = \emptyset$.

Corollary 2. *([9]) Let θ be an antimorphic involution on Σ^* and let $u \in \Sigma^+$. Then $u \in \mathrm{D}_\theta(1)$ if and only if $u^+ \subseteq \mathrm{D}_\theta(1)$.*

In contrast to Corollary 2, it is not always the case that, given a θ-pseudoknot-unbordered word u, the word u^k remains θ-pseudoknot-unbordered for any k, that is, in general we cannot say that $u \in \mathrm{K}_\theta(1)$ if and only if $u^+ \subseteq \mathrm{K}_\theta(1)$. See the next example.

Example 7. Let θ be an antimorphic involution on Σ^*, and let $a, b \in \Sigma$ such that $\theta(a) = b$ and $\theta(b) = a$. Since θ does not equal the identity, we can always find such letters a and b in Σ.

Let $u = aabbbbaba$. Although $u \in \mathrm{K}_\theta(1)$, uu is θ-pseudoknot-bordered for $x = aabbb$ and $y = babaa$. In fact, $uu = xyabbbbaba = aabbbbab\theta(x)\theta(y)$. □

Nevertheless, when θ is an antimorphic involution, we can give a characterization of such counterexamples that takes into account the relative length of the θ-pseudoknot-borders.

Proposition 2. *Let θ be an antimorphic involution on Σ^*. Then for a θ-pseudoknot-unbordered word $u \in \mathrm{K}_\theta(1)$, if there exists $k \geq 2$ such that u^k has a nonempty θ-pseudoknot-border w, then $|u| < |w| < \frac{4}{3}|u|$ holds.*

Proof. Suppose for some $k \geq 2$, there were a $w \in \mathrm{L}^\theta_{\mathrm{cd}}(u^k)$ such that either $|w| \leq |u|$ or $\frac{4}{3}|u| \leq |w|$ hold. If $|w| \leq |u|$, then this w leads us to the contradiction immediately. Next we consider the case $\frac{4}{3}|u| \leq |w| < 2|u|$. Then $w \in \mathrm{L}^\theta_{\mathrm{cd}}(u^k)$ implies $w \in \mathrm{L}^\theta_{\mathrm{cd}}(u^2)$. In other words, there exists a decomposition $w = xy$ such that $uu = xy\alpha = \beta\theta(x)\theta(y)$ for some $\alpha, \beta \in \Sigma^+$. Since $|w| \geq \frac{4}{3}|u|$, we have $xy = uu_p$ and $\theta(x)\theta(y) = u_s u$, where $u_p \in \mathrm{Pref}(u)$, and $u_s \in \mathrm{Suff}(u)$. Now we have the following two cases: (1) $|x| \geq |u|$ or $|y| \geq |u|$ holds, or (2) $|x| < |u|$ and $|y| < |u|$ hold.

In the first case, for reasons of symmetry, we only have to consider the case $|x| \geq |u|$. Since $\theta(x)\theta(y) = u_s u$, we can write $\theta(x) = u_s u'_p$ for some $u'_p \in \mathrm{Pref}(u)$. Let $u = u'_p u'_s$, and we can easily check that $u'_s \in \mathrm{Suff}(u_s)$. Therefore, $u'_s u'_p \in \mathrm{Suff}(\theta(x))$, which equals $\theta(u'_p)\theta(u'_s) \in \mathrm{Pref}(x)$. This means that $\theta(u'_p)\theta(u'_s) = u$ because u and $\theta(u'_p)\theta(u'_s)$ are prefixes of x and they have equal lengths. Since $u = u'_p u'_s$, we coclude that both u'_p and u'_s are θ-palindromes. The application of Lemmata 3 and 4 leads now to a contradiction.

Next we consider the second case. This figure shows $xy = uu_p$ and $\theta(x)\theta(y) = u_s u$. Since both x and y are shorter than u, these equations imply that $u = xu'_s = u'_p\theta(y)$, where $u'_p \in \mathrm{Pref}(u)$ and $u'_s \in \mathrm{Suff}(u)$. Comparing this equation with $xy = uu_p$ we derive $y = u'_s u_p$, and hence $u = u'_p\theta(u_p)\theta(u'_s)$. This result, together with $u = xu'_s$, implies that u'_s is a θ-palindrome and $x = u'_p\theta(u_p)$. Substituting this x and $u = u'_p\theta(y)$ into $\theta(x)\theta(y) = u_s u$ gives $u_p\theta(u'_p)\theta(y) = u_s u'_p\theta(y)$, which means that $u_p = u_s$ and u'_p is a θ-palindrome.

Let us bring now into the picture the original condition $\frac{4}{3}|u| \leq |w| < 2|u|$. Since $|w| = |u| + |u_p|$, $\frac{4}{3}|u| \leq |w|$ means $\frac{1}{3}|u| \leq |u_p|$. Hence, $|xy| = |uu_p| \leq 4|u_p|$. This implies that either $|x| \leq 2|u_p|$ or $|y| \leq 2|u_p|$ holds. We assume the former case holds. Then $\theta(x) = u_s u'_p$ implies $|u'_p| \leq |u_s|$ because $|\theta(x)| = |x| \leq 2|u_p| = 2|u_s|$. Let $u_s = u_1 u_2$ such that $|u_1| = |u'_p|$. Note that $u_s \in \mathrm{Pref}(x)$ because $u_p, x \in \mathrm{Pref}(u)$, $|u_s| < |x|$, and $u_p = u_s$. Comparing $u_s = u_1 u_2$ with $x = \theta(u_1 u_2 u'_p)$ based on $u_s \in \mathrm{Pref}(x)$ results in $u_2 = \theta(u_2)$ and $u_1 = \theta(u'_p)$, which in turn derives $u_1 = \theta(u_1)$ because $u'_p = \theta(u'_p)$. Now Lemmata 3 and 4 lead to a contradiction because u contains the concatenation of two θ-palindromes u_1 and u_2 as its prefix u_p and suffix u_s.

In the case $|w| \geq 2|u|$, either $|x| \geq |u|$ or $|y| \geq |u|$ holds. Thus, we get a contradiction in a similar way as above. □

Actually, in Example 7, the θ-pseudoknot-border xy of u^2 satisfies $|u| < |xy| < \frac{4}{3}|u|$.

Corollary 3. *Let θ be an antimorphic involution on Σ^*. For a word $u \in \mathrm{K}_\theta(1)$, $u^+ \not\subseteq \mathrm{K}_\theta(1)$ if and only if $u^2 \notin \mathrm{K}_\theta(1)$.*

In what follows, we give a characterization of such a θ-pseudoknot-unbordered word u with the property that u^2 is not included in $\mathrm{K}_\theta(1)$.

Lemma 6. *Let θ be an antimorphic involution on Σ^*, and let u be a θ-pseudoknot-unbordered word, i.e., $u \in \mathrm{K}_\theta(1)$. Then u^2 has a θ-pseudoknot-border if and only if $u = u_p\alpha\theta(u_p)\beta u_p$ for some $u_p, \alpha, \beta \in \Sigma^+$ such that $u_p\alpha, \beta u_p$ are θ-palindromes.*

Lemma 7. *Let θ be an antimorphic involution on Σ^* and u be a θ-pseudoknot-unbordered word, i.e., $u \in \mathrm{K}_\theta(1)$. If u^2 is θ-pseudoknot-bordered, then u is primitive.*

As a contraposition of this lemma, the following corollary holds.

Corollary 4. *If $u \in \mathrm{K}_\theta(1)$ and it is not primitive, then u^2 is θ-pseudoknot-unbordered, i.e., $u^2 \in \mathrm{K}_\theta(1)$. This further implies that $u^+ \subseteq \mathrm{K}_\theta(1)$.*

To complete this discussion, we note that there exists an antimorphic involution θ and a non-primitive word u such that u^k is not θ-pseudoknot-bordered for any $k > 1$.

Example 8. Let $u = abaaabaa$, which is clearly not primitive, and θ be an antimorphic involution such that $\theta(a) = b$ and vice versa. It is easy to see that neither u nor uu are θ-pseudoknot-bordered. □

Lastly, as the next result shows, given a θ-pseudoknot-unbordered word u, if u^2 is θ-pseudoknot-bordered then u and any θ-pseudoknot-border of u^2 are primitive.

Theorem 1. *Let θ be an antimorphic involution on Σ^*, and $u \in \mathrm{K}_\theta(1)$ satisfying $u^+ \not\subseteq \mathrm{K}_\theta(1)$. Then any θ-pseudoknot-border of u^2 is primitive.*

The rest of this section will show that for a word $u \in \mathrm{K}_\theta(1)$, the factorization of a θ-pseudoknot-border $w \in \mathrm{L}^\theta_{\mathrm{cd}}(u^2)$ into x and y is unique. In other words, $w = xy = x'y'$ such that $u^2 = xy\alpha = \beta\theta(x)\theta(y)$ and $u^2 = x'y'\alpha' = \beta'\theta(x')\theta(y')$ mean $x = x'$ and $y = y'$. Note that $x \neq y$ because if they were equal, this border xy would not be primitive, which conflicts with Theorem 1.

Lemma 8. *Let θ be an antimorphic involution on Σ^*, and $w \in \Sigma^*$. For $xy \in \mathrm{L}^\theta_{\mathrm{cd}}(w)$ such that $x \neq y$, $xy = uv = vu$ for some different words $u, v \in \Sigma^+$ if and only if w has a different θ-pseudoknot-border $x'y'$ of the same length as xy, i.e., $x'y' = xy$ but $|x'| \neq |x|$.*

Proposition 3. *Let θ be an antimorphic involution on Σ^*, $w \in \Sigma^*$, and $u \in \mathrm{K}_\theta(1)$. If w is a θ-pseudoknot-border of u^2, then the factorization of w into x and y such that $u^2 = xy\alpha = \beta\theta(x)\theta(y)$ for some $\alpha, \beta \in \Sigma^*$ is unique.*

References

1. Adleman, L.: Molecular computation of solutions to combinatorial problems. Science 266, 1021–1024 (1994)
2. Daley, M., Kari, L.: DNA computing: Models and Implementations. Comments on Theoretical Biology 7(3), 177–198 (2002)
3. McCaskill, J.S.: The equilibrium partition function and base pair binding probability for RNA secondary structure. Biopolymers 29, 1105–1119 (1990)
4. Andronescu, M., Dees, D., Slaybaugh, L., Zhao, Y., Condon, A., Cohen, B., Skiena, S.: Algorithms for testing that sets of DNA words concatenate without secondary structure. In: Hagiya, M., Ohuchi, A. (eds.) DNA 2002. LNCS, vol. 2568, pp. 182–195. Springer, Heidelberg (2003)
5. Condon, A.E.: Problems on RNA secondary structure prediction and design. In: ICALP 2003. LNCS, vol. 2719, pp. 22–32. Springer, Heidelberg (2003)
6. Kari, L., Kitto, R., Thierrin, G.: Codes, involutions and DNA encodings. In: Brauer, W., Ehrig, H., Karhumäki, J., Salomaa, A. (eds.) FNC 2002. LNCS, vol. 2300, pp. 376–393. Springer, Heidelberg (2002)
7. Kari, L., Konstantinidis, S., Losseva, E., Wozniak, G.: Sticky-free and overhang-free DNA languages. Acta Informatica 40, 119–157 (2003)
8. Kari, L., Konstantinidis, S., Losseva, E., Sosík, P., Thierrin, G.: Hairpin structures in DNA words. In: Carbone, A., Pierce, N.A. (eds.) DNA 2005. LNCS, vol. 3892, pp. 158–170. Springer, Heidelberg (2006)
9. Kari, L., Mahalingam, K.: Involutively bordered words. International Journal of Foundations of Computer Science (2007)
10. Kobayashi, S.: Testing structure freeness of regular sets of biomolecular sequences (extended abstract). In: Ferretti, C., Mauri, G., Zandron, C. (eds.) DNA 2004. LNCS, vol. 3384, pp. 192–201. Springer, Heidelberg (2005)
11. Kari, L., Seki, S.: On pseudoknot words and their properties (submitted)
12. Jones, S.-G., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A.: Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acid Research 33, 121–124 (2005)
13. Grätzer, G.: Universal Algebra. Van Nostrand Princeton, NJ (1968)
14. Rozenberg, G., Salomaa, A. (eds.): Handbook of Formal Languages. Springer, Berlin (1997)
15. Lentin, A., Schützenberger, M.P.: A combinatorial problem in the theory of free monoids. In: Proceedings of Combinatorial Mathematics and its Applications, April 10-14, pp. 128–144 (1967)
16. Ehrenfeucht, A., Silberger, D.: Periodicity and unbordered segments of words. Discrete Mathematics 26(2), 101–109 (1979)
17. Jonoska, N., Mahalingam, K., Chen, J.: Involution codes: with application to DNA coded languages. Natural Computing 4(2), 141–162 (2005)