# Watson-Crick bordered words and their syntactic monoid

Lila Kari and Kalpana Mahalingam

University of Western Ontario,
Department of Computer Science,
London, ON, Canada N6A 5B7
lila, kalpana@csd.uwo.ca

**Abstract.** *DNA strands that, mathematically speaking, are finite strings over the alphabet $\{A, G, C, T\}$ are used in DNA computing to encode information. Due to the fact that A is Watson-Crick complementary to T and G to C, DNA single strands that are Watson-Crick complementary can bind to each other or to themselves in either intended or unintended ways. One of the structures that is usually undesirable for biocomputation, since it makes the affected DNA string unavailable for future interactions, is the hairpin: If some subsequences of a DNA single string are complementary to each other, the string will bind to itself forming a hairpin-like structure. This paper studies a mathematical formalization of a particular case of hairpins, the Watson-Crick bordered words. A Watson-Crick bordered word is a word with the property that it has a prefix that is Watson-Crick complementary to its suffix. We namely study algebraic properties of Watson-Crick bordered and unbordered words. We also give a complete characterization of the syntactic monoid of the language consisting of all Watson-Crick bordered words over a given alphabet. Our results hold for the more general case where the Watson-Crick complement function is replaced by an arbitrary antimorphic involution.*

## 1 Introduction

The subject of this paper, Watson-Crick (WK) bordered words, is motivated by the practical requirements of DNA computing experiments. DNA strands can be viewed as finite strings over the alphabet $\{A, G, C, T\}$ and are used in DNA computing to encode information. Since $A$ is Watson-Crick complementary to $T$ and $G$ to $C$, DNA single strands that are WK complementary can bind to each other or to themselves in either intended or unintended ways. One of these undesirable DNA secondary structures, the *hairpin*, is formed when the suffix of a DNA single strand is WK complementary to the prefix of the same DNA strand. A word with this property is called Watson-Crick bordered. Experimentally, DNA strands that are Watson-Crick bordered are to be avoided when encoding data on DNA strands, since the hairpin structures they form make them unavailable for biocomputations. Theoretically, Watson-Crick bordered words generalize the classical definition of a bordered word: A bordered word is one with the property that it has a prefix that equals its suffix, [20], [18].

If in a Watson-Crick bordered word over the DNA alphabet the prefix and its WK complementary suffix do not overlap, then the strand forms a hairpin structure such as the one shown in Fig 1. If, on the other hand, the prefix of such a word and the WK complement of one of its suffixes overlap, the DNA strand could bind with another copy of itself as shown in Fig 2. Both such bindings are potentially undesirable for DNA computing experiments and this paper investigates words that could potentially interact this way. Algebraic properties of other types of languages that avoid DNA sequences undesirable for DNA based computations, such as sticky-free languages, overhang-free languages and hairpin-free languages, have

been extensively studied in [2, 3, 5, 8, 9]. The notion of Watson-Crick bordered words was formalized and its coding properties as well as relations between Watson-Crick bordered words and other types of codes have been discussed in [11]. Certain algebraic properties of involution bordered words were discussed in [11]. In this paper we study the algebraic properties of the set of all Watson-Crick bordered words through their syntactic monoid.


**Fig. 1.** If a word $u$ is Watson-Crick bordered and its WK borders do not overlap, the word $u$ may stick to itself forming a simple hairpin loop, as shown above.

The reason for our choice of method of investigation is that the syntactic monoid approach to the study of a language has proved to be very fruitful in other cases. Algebraic characterizations of many classes of codes through their syntactic monoid have been extensively studied [6, 14–16, 19]. In [6], the author formulated a general characterization method of the syntactic monoid which applies to all classes of codes that can be defined in a certain way and hence results analogous to those of [16] can be obtained for a large variety of classes of codes. For more details on codes the reader is referred to [1, 7, 18].

More recently, in [10] we have discussed the syntactic monoid properties of the set of all hairpin-free words. In this paper we use these methods to study the algebraic properties of the set of all involution-bordered words. Throughout the paper we concentrate on an antimorphic involution $\theta$ such that $\theta(a) \neq a$ for all $a \in \Sigma$. Such a function is arguably an accurate mathematical formalization of the Watson-Crick DNA strand complementarity as it features its main properties: the fact that the WK complement of a DNA strand is the reverse (antimorphism property) complement (involution property) of the original strand. (An involution is a function $\theta$ such that $\theta^2$ equals the identity.)

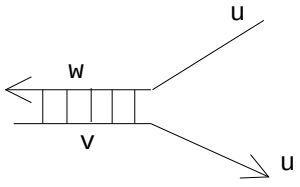


**Fig. 2.** If a word $u$ is Watson-Crick bordered and its WK borders overlap, the word $u$ may stick to another copy of itself as shown above. (Usually, in a DNA computing experiment, each DNA strand is present in hundreds or millions of copies in the solution.)

The paper is organized as follows: Section 2 reviews basic definitions. It is easy to see that, for an antimorphic involution, the set of all involution-bordered words is a proper subset of the set of all hairpins as studied in [10]. (Note that neither this

inclusion nor its reverse hold if we consider the set of general hairpins of a given length $k$). In [10] we showed that the elements of the syntactic monoid of the language of all hairpin-free words are idempotents and the monoid is commutative. In this paper (Section 3) we obtain a different result for involution-bordered word sets: We now show that, while all the elements of the syntactic monoid of the language of all involution-bordered words over a given alphabet are idempotents, the monoid is not commutative. We also observe that similarly to the case of the hairpin-free words, the language of all involution-bordered words is locally testable. Proposition 5 and 6 parallel results in [10] by giving a necessary and sufficient condition for a finite monoid to be the syntactic monoid of the set of all involution-bordered words over a given finite alphabet. In Section 4, we discuss the Green's relations for the set of all involution-bordered words. In contrast to the case of the set of all hairpin-free words, it turns out that the Green's relations are not trivial for the set of all involution-bordered words.

## 2 Definitions and basic concepts

In this section we review some basic notions. An alphabet set $\Sigma$ is a finite non-empty set of symbols. A word $u$ over $\Sigma$ is a finite sequence of symbols in $\Sigma$. We denote by $\Sigma^*$ the set of all words over $\Sigma$, and by $\Sigma^+$ the set of all non empty words over $\Sigma$. The empty word is denoted by $\lambda$. We note that with the concatenation operation on words, $\Sigma^*$ is the free monoid and $\Sigma^+$ is the free semigroup generated by $\Sigma$. The length of a word $u = a_1...a_n$ is $n$ for all $a_i \in \Sigma$ and is denoted by $|u|$. A language over $\Sigma$ is an arbitrary subset of $\Sigma^*$. A mapping $\theta : \Sigma^* \mapsto \Sigma^*$ is called a morphism (antimorphism) of $\Sigma^*$ if $\theta(uv) = \theta(u)\theta(v)$ (respectively $\theta(uv) = \theta(v)\theta(u)$) for all $u, v \in \Sigma^*$. An involution map $\theta$ is such that $\theta^2$ equals identity.

Bordered words were initially called "overlapping words" and unbordered words were called as "non-overlapping words", [18]. For properties of bordered and unbordered words we refer the reader to [20], [18]. In [11], we extended the concept of bordered words to involution-bordered words and studied some of their algebraic properties. We now recall some definitions defined and used in [11].

**Definition 1.** *Let $\theta$ be either a morphic or an antimorphic involution on $\Sigma^*$.*

1. *A word $u \in \Sigma^+$ is said to be $\theta$-bordered if there exists $v \in \Sigma^+$ such that $u = vx = y\theta(v)$ for some $x, y \in \Sigma^+$.*
2. *A non-empty word which is not $\theta$-bordered is called $\theta$-unbordered.*

**Lemma 1** *Let $\theta$ be either morphic or an antimorphic involution.*

1. *A $\theta$-bordered word $x \in \Sigma^+$ has length greater than or equal to 2.*
2. *For all $a \in \Sigma$, $a$ is $\theta$-unbordered.*
3. *For all $a \in \Sigma$ such that $a \neq \theta(a)$, $a^n$ is $\theta$-unbordered for all $n \geq 1$.*

In case $\theta$ is the Watson-Crick involution a $\theta$-bordered word will be called Watson-Crick bordered, and a $\theta$-unbordered word will be called Watson-Crick unbordered. Figures 1 and 2 illustrate some undesirable interactions that can result if a DNA string is Watson-Crick bordered.

We recall that a language or a set $X \subseteq \Sigma^*$ is said to be dense if for all $u \in \Sigma^*$, $X \cap \Sigma^* u \Sigma^* \neq \emptyset$. The following lemma was proved in [11].

**Lemma 2** *Let $\theta$ be an antimorphic involution. Let $L$ be the set of all $\theta$-bordered words over $\Sigma^*$. Then*

1. *$L$ is regular.*
2. *$L$ is a dense set.*

## 3 The syntactic monoid of the set of all Watson-Crick bordered words

In the theory of codes, two types of syntactic monoids are usually considered, the syntactic monoid of the code itself and the syntactic monoid of the Kleene star of the code. In this section we concentrate on the characterizations of syntactic monoid of the set of all $\theta$-bordered words, when $\theta$ is an antimorphic involution such that $\theta(a) \neq a$ for all $a \in \Sigma$. Necessary and sufficient conditions for a monoid to be the syntactic monoid of the set of all $\theta$-bordered words are also discussed. We first review some basic concepts.

Let $L$ be a language such that $L \subseteq \Sigma^+$. We define the context, right context and left context of a word $w \in \Sigma^*$ in $L$ as follows:

- $C_L(w) = \{(u, v) : uwv \in L, u, v \in \Sigma^*\}$.
- $\mathcal{R}_L(w) = \{u \in \Sigma^* : wu \in L\}$.
- $\mathcal{L}_L(w) = \{u \in \Sigma^* : uw \in L\}$.

$C_L(w)$, $\mathcal{R}_L(w)$ and $\mathcal{L}_L(w)$ are called the context, right context and left context of $w$ in $L$ respectively. Also note that $Sub(L) = \{x : pxq \in L, p, q \in \Sigma^*\}$ is the set of all subwords of $L$. Recall that

**Definition 2.** *Let $L$ be a language such that $L \subseteq \Sigma^+$.*

1. *The syntactic congruence of $L \subseteq \Sigma^+$ is denoted by $P_L$ and is defined by $u \equiv v(P_L)$ iff $C_L(u) = C_L(v)$.*
2. *The syntactic monoid of $L$ is the quotient monoid $M(L) = \Sigma^*/P_L$ with the operation $[x][y] = [xy]$, where for $x \in \Sigma^*$, $[x]$ denotes the $P_L$ equivalence class of $x$.*

Let $W(L) = \{x \in \Sigma^* : C_L(x) = \emptyset\}$, i.e., $x \in W(L)$ iff $x \notin Sub(L)$. $W(L)$ is called the residue of $L$.

Note that if $W(L) \neq \emptyset$ then $W(L)$ represents a class for $P_L$ and is the zero of $M(L)$.

Note that for a regular language $L$, $M(L)$ is the transition monoid (see [17]) of the minimal deterministic finite automaton (see [1, 17]) of $L$. The above definition of the syntactic congruence $P_L$ can be defined for an arbitrary subset $L$ of any semigroup $S$. If the syntactic congruence is the equality relation then we call the set $L$ to be a disjunctive subset of $S$. If $L = \{x\}$ for some $x \in \Sigma^*$ and if $P_L$ is the equality relation then we say that $x$ is a disjunctive element of $S$. For more on syntactic monoid we refer the reader to [1, 12, 17].

It is a well known fact that $L$ is a regular language if and only if $M(L)$ is finite (see [12, 17]). For any set $L$ and its syntactic monoid $M(L)$, $\eta : \Sigma^* \to M(L)$ is the natural surjective syntactic morphism defined by $x \to [x]$. Note that for any $L$, $L$ is a union of $P_L$ classes.

We denote by $B_{\theta,\Sigma}$ the set of all $\theta$-bordered words over $\Sigma^*$, with $\theta$ an antimorphic involution and $\theta(a) \neq a$ for all $a \in \Sigma$. In the remainder of the paper, if the alphabet $\Sigma$ is clear from the context, we will denote the set of all $\theta$-bordered words over $\Sigma$ simply by $B_\theta$.

It was shown in [11] that $B_\theta$ is regular and hence $Syn(B_\theta)$ is finite. In the following lemma we show that the residue of $B_\theta$ is the empty set.

**Lemma 3** *The residue of $B_\theta$ is the empty set, i.e., $W(B_\theta) = \emptyset$.*

*Proof.* Follows from the fact that $B_\theta$ is dense, see Lemma 2. □

In the following proposition we show that every non zero element of $Syn(B_\theta)$ is idempotent.

**Proposition 1** *For every $u \in \Sigma^*$, we have $u\ P_{B_\theta}\ u^2$.*

*Proof.* The congruence $P_{B_\theta}$ is equivalent to the congruence $P_{\overline{B_\theta}}$ associated to the complement $\overline{B_\theta}$ of $B_\theta$. Hence we have to show that $u\ P_{\overline{B_\theta}}\ u^2$, i.e., $xuy \in \overline{B_\theta}$ iff $xu^2y \in \overline{B_\theta}$. Assume that $xuy \in \overline{B_\theta}$. Suppose that $xu^2y \in B_\theta$, then there exists $a \in \Sigma$ such that $xu^2y = av\theta(a)$ for some $v \in \Sigma^*$. We have the following cases:

1. If $x = ax_1$ and $y = y_1\theta(a)$ then $xuy = ax_1uy_1\theta(a)$, a contradiction since $xuy \in \overline{B_\theta}$.
2. If $x = \lambda$, the empty word, then $u^2y = av\theta(a)$ which implies $u = av_1$ and $y = y_1\theta(a)$ and hence $xuy = av_1y_1\theta(a)$, again a contradiction. The case when $y = \lambda$ is similar.
3. If both $x$ and $y$ are empty, i.e., $x = y = \lambda$, then $u^2 = av\theta(a)$. If $v = \lambda$, then $u = a = \theta(a)$ a contradiction to our assumption that $a \neq \theta(a)$ for all $a \in \Sigma$. Thus $v \neq \lambda$ and $u = av_1 = v_2\theta(a)$ a contradiction since $xuy = u \in \overline{B_\theta}$.

Hence $xu^2y \in \overline{B_\theta}$. Conversely, assume that $xu^2y \in \overline{B_\theta}$. Suppose $xuy \in B_\theta$, then there exists $a \in \Sigma$ such that $xuy = av\theta(a)$ for some $v \in \Sigma^*$. We have the following cases:

1. If $x = ax_1$ and $y = y_1\theta(a)$ then $xu^2y = ax_1u^2y_1\theta(a)$, a contradiction since $xu^2y \in \overline{B_\theta}$.
2. If $x = \lambda$, the empty word, then $uy = av\theta(a)$ which implies $u = av_1$ and $y = y_1\theta(a)$ and hence $xu^2y = av_1uy_1\theta(a)$, again a contradiction. The case when $y = \lambda$ is similar.
3. If both $x$ and $y$ are empty, i.e., $x = y = \lambda$, then $u = av\theta(a)$ which implies that $xu^2y = u^2 = av\theta(a)av\theta(a)$ again a contradiction since $xu^2y \in \overline{B_\theta}$.

Thus $xuy \in \overline{B_\theta}$ iff $xu^2y \in \overline{B_\theta}$ and hence $uP_{\overline{B_\theta}}u^2$ for all $u \in \Sigma^*$. □

**Corollary 1** *The elements of the syntactic monoid of $B_\theta$ are idempotent elements.*

*Proof.* The fact that $uP_{\overline{B_\theta}}u^2$ for any $u \in \Sigma^*$ implies that $U = U^2$ for the class $U$ containing $u$. □

If $\theta$ is a mapping of $\Sigma^*$ into $\Sigma^*$, a congruence $R$ is said to be $\theta$-compatible if $uRv$ implies $\theta(u)R\theta(v)$. If such is the case, then the mapping $\theta$ on $\Sigma^*$ can be extended to a mapping of the quotient-monoid $S = \Sigma^*/R$ in the following way. Let $U$ be the class mod $R$ containing the word $u$. Define $\theta(U)$ to be the class of $R$ containing $\theta(u)$. This mapping is well defined, i.e., it does not depend on the choice of the representative $u$ of the class $U$. Indeed if $u' \in U$, then, $R$ being $\theta$-compatible, we have $\theta(u)R\theta(u')$ and hence $\theta(u') \in \theta(U)$.

**Proposition 2** *The syntactic congruence $P_{B_\theta}$ is $\theta$-compatible.*

*Proof.* To show that $P_{B_\theta}$ is $\theta$-compatible, we have to show that $uP_{B_\theta}v$ implies $\theta(u)P_{B_\theta}\theta(v)$, i.e., $C_{B_\theta}(u) = C_{B_\theta}(v)$ implies $C_{B_\theta}(\theta(u)) = C_{B_\theta}(\theta(v))$. Let $uP_{B_\theta}v$ and let $(x, y) \in C_{B_\theta}(\theta(u))$, then $x\theta(u)y \in B_\theta$ which implies that $\theta(x\theta(u)y) \in \theta(B_\theta)$. Thus $\theta(y)\theta(\theta(u))\theta(x) \in \theta(B_\theta)$, .i.e., $\theta(y)u\theta(x) \in \theta(B_\theta)$. Since $B_\theta$ is $\theta$ stable, $\theta(B_\theta) \subseteq B_\theta$ and thus $\theta(y)u\theta(x) \in B_\theta$ iff $\theta(y)v\theta(x) \in B_\theta$ since $uP_{B_\theta}v$. Therefore $\theta(\theta(y)v\theta(x)) \in \theta(B_\theta) \subseteq B_\theta$ and therefore $x\theta(v)y \in B_\theta$ which implies that $(x, y) \in C_{B_\theta}(\theta(v))$. Similarly we can show that $C_{B_\theta}(\theta(v)) \subseteq C_{B_\theta}(\theta(u))$. Thus $P_{B_\theta}$ is $\theta$-compatible. □

Recall that a semigroup in general is a set equipped with an internal associative operation which is usually written in a multiplicative form. A monoid is a semigroup with an identity element (usually denoted by $e$). If $S$ is a semigroup, $S^1$ denotes the monoid equal to $S$ if $S$ has an identity element and to $S \cup \{e\}$ otherwise. In the latter case, the multiplication on $S$ is extended by setting $s.e = e.s = s$ for all $s \in S$. Let $e \in S$ be an idempotent of S. Then the set $eSe = \{ese : s \in S\}$ is a subsemigroup of $S$, called the local subsemigroup associated with $e$. This semigroup is in fact a monoid, since $e$ is an identity in $eSe$. We also recall that a semigroup $S$ is called locally trivial if for all $s \in S$ and for all idempotents $e \in S$, we have $ese = e$. We recall the following result.

**Proposition 3** *[17] Let $S$ be a non empty semigroup. The following are equivalent.*

1. *$S$ is locally trivial.*
2. *The set of all idempotents is the minimal ideal of $S$.*
3. *We have $esf = ef$ for all $s \in S$ and for all idempotents $e, f \in S$.*

Since for all $e \in Syn(B_\theta)$, $e$ is an idempotent, we have the following observations. Let $S = Syn(B_\theta) \setminus \{1\}$, then

- $S$ is aperiodic, i.e., for all $e \in S$, there exists $n$ such that $e^n = e^{n+1}$.
- $S$ is regular, i.e., for all $e \in S$, $e$ is regular, i.e., there exists $s \in S$ such that $ese = e$.

**Lemma 4** *For all $[ab] \in Syn(B_\theta)$, such that $a, b \in \Sigma$, $[ab]$ as a set is equal to the set of all words that begin with $a$ and end with $b$.*

*Proof.* We first prove for the case when $a \neq b$. Clearly $ab \in [ab]$. Let $u \in \Sigma^*$ be such that $aub \notin [ab]$. Then there exists $x, y \in \Sigma^*$ such that $xaby \in B_\theta$ and $xauby \notin B_\theta$. Note that $xaby \in B_\theta$ implies that $xaby = cp\theta(c)$ for some $c \in \Sigma$ and $p \in \Sigma^*$. Then $xauby = cq\theta(c)$ which implies that $xauby \in B_\theta$ a contradiction. Hence $aub \in [ab]$ for all $u \in \Sigma^*$.
If $a = b$, then clearly we have $aa \in [aa]$ and for all $u \in \Sigma^*$, $aua \in [aa]$. Suppose $a \notin [aa]$ then there exists $x, y \in \Sigma^*$ such that $xaay \in B_\theta$ and $xay \notin B_\theta$. Note that $xaay \in B_\theta$ implies that $xaay = cp\theta(c)$ for some $c \in \Sigma$ and $p \in \Sigma^*$. If both $x$ and $y$ are non empty, then $xay = cq\theta(c)$ for some $c \in \Sigma$ and $q \in \Sigma^*$, which implies that $xay \in B_\theta$, which is a contradiction. If $x = \lambda$ and $y \in \Sigma^+$ then $aay = cp\theta(c)$ which implies $a = c$ and $y = y_1\theta(c)$ and hence $xay = ay = cy_1\theta(c)$ which implies that $xay \in B_\theta$, a contradiction. The case when $x \in \Sigma^+$ and $y = \lambda$ is similar. If $x = y = \lambda$, then $aa = cp\theta(c)$ which implies that $a = c = \theta(c)$ a contradiction to our assumption, since for all $a \in \Sigma$, $\theta(a) \neq a$. Hence $a \in [aa]$. Thus for all $a, b \in \Sigma$, and for all $[ab] \in Syn(B_\theta)$, $[ab]$ as a set is the set of all words that begin with $a$ and end with $b$. □

Recall that a language $L$ is said to be $n$-locally testable if whenever $u$ and $v$ have the same factors of length at most $n$ and the same prefix and suffix of length $n-1$ and $u \in L$ then $v \in L$. The language $L$ is locally testable if it is $n$-locally testable for some $n \in I\!N$.

We also recall a characterization of the syntactic semigroup of locally testable languages which states that (Proposition 2.1 in [13]) a recognizable subset (A language is called recognizable if there exists an algorithm that accepts a given string if and only if the string belongs to that language) $L$ of $\Sigma^+$ is locally testable iff for all idempotents $g \in Syn(L)$, $gSyn(L)g$ is a semi lattice. We use this characterization and the above proposition to show that $B_\theta$ is locally testable.

**Corollary 2** *$B_\theta$ is locally testable.*

*Proof.* We need to show that for all $e \in Syn(B_\theta)$, $eSyn(b)e$ is a semilattice. Note that from Lemma 4, for all $e, s \in Syn(B_\theta)$, $ese = e$ and hence $eSyn(B_\theta)e = \{e\}$. Since $e$ is an idempotent and $\{e\}$ is commutative, $eSyn(B_\theta)e = \{e\}$ is a semilattice. Thus $B_\theta$ is locally testable.

**Corollary 3** *$S = Syn(B_\theta) \setminus \{1\}$ is locally trivial.*

*Proof.* For all $e \in S$, $e$ is an idempotent. We need to show that $ese = e$ for all $e, s \in S$. Let $e = [ab]$ for some $a, b \in \Sigma$ and let $s = [s_1]$ for some $s_1 \in \Sigma^+$. Then $ese = [ab][s_1][ab] = [abs_1ab] = [ab] = e$. Hence $S$ is locally trivial. □

**Corollary 4** *$S$ is the minimal ideal of $S$ and for all $e, s, f \in S$, $esf = ef$.*

*Proof.* Follows from the fact that $S$ is locally trivial and all elements of $S$ are idempotents and from Proposition 3. □

**Corollary 5** *For all $e, f, g \in Syn(B_\theta)$, if $eg = fg$ and $ge = gf$ then $e = f$.*

*Proof.* Given that $eg = fg$ and $ge = gf$. Then $eg.ge = fg.gf$ which implies that $ege = fgf$ since for all $e \in Syn(B_\theta)$, $e$ is an idempotent. Thus from Corollary 4, $ege = e^2 = e = fgf = f^2 = f$ which implies that $e = f$.

**Corollary 6** *$Syn(B_\theta)$ is a simple semigroup.*

*Proof.* Since $\emptyset$ and $S = Syn(B_\theta)$ are the only ideals of $Syn(B_\theta)$, $S$ is simple.

In the next proposition we show that for all $e, f \in S$, $e$ and $f$ are conjugates, i.e., $e = uv$ and $f = vu$ for some $u, v \in S$.

**Proposition 4** *For all $e, f \in S$, $e$ and $f$ are conjugates.*

*Proof.* Let $e, f \in S$ such that $e = [ab]$ and $f = [cd]$ for some $a, b, c, d \in \Sigma$. Then $e = [ab] = [adcb] = [ad][cb]$ and $f = [cd] = [cbad] = [cb][ad]$ which implies that $e$ and $f$ are conjugates. □

**Lemma 5** *$P_{B_\theta}$ class of 1 is trivial.*

*Proof.* Suppose not, let $u \equiv 1(P_{B_\theta})$ for some $u \in \Sigma^+$. Then for any $v \in B_\theta$, $uv \equiv v(P_{B_\theta})$ and $vu \equiv v(P_{B_\theta})$. Since $v \in B_\theta$, $uv, vu \in B_\theta$. Also, $v, uv, vu \in [ab]$ for some $a, b \in \Sigma$ with $\theta(a) = b$. Thus $v = axb$, $uv = ayb$ and $vu = azb$ for some $x, y, z \in \Sigma^*$. Then $u = arb$ which implies that $u \in [ab]$ and hence $1 \in [ab]$ a contradiction since $1 \notin B_\theta$. Thus $P_{B_\theta}$ class of 1 is trivial.

In the following results, using the notion of the syntactic monoid, similar to Proposition 17, 18 in [10], we establish an algebraic connection between the language $B_\theta$ of the bordered words relatively to an antimorphic involution $\theta$ over a finite alphabet $\Sigma$ and a certain class of finite monoids.

**Proposition 5** *Let $Syn(B_\theta)$ be the syntactic monoid of $B_\theta$. Then:*

1. *$Syn(B_\theta)$ is a finite monoid which has no zero and every element of $Syn(B_\theta)$ is idempotent.*

2. *There exists an antimorphic involution $\psi$ such that the set $Syn(B_\theta)$ is stable under $\psi$.*
3. *$Syn(B_\theta)$ has two non empty disjunctive sets $D_1$ and $D_2$ such that $Syn(B_\theta) = D_1 \cup D_2$ and $D_1 \cap D_2 = \emptyset$, where $D_1 = \{[x] \in Syn(B_\theta) \setminus \{1\} : \psi([x]) = [x]\}$.*

*Proof.* 1. The regularity of the language $B_\theta$ implies the finiteness of its syntactic monoid $Syn(B_\theta)$. Since $B_\theta$ is dense, $Syn(B_\theta)$ has no zero. The last part follows from Corollary 1.

2. Since the syntactic congruence $P_{B_\theta}$ is $\theta$-compatible, an antimorphic involution $\psi$ can be defined on $Syn(B_\theta)$ in the following way. Let $U$ be an element of $Syn(B_\theta)$, i.e., $U$ is a class of $P_{B_\theta}$, and define $\psi(U)$ to be the class containing the element $\theta(u)$, where $u \in U$. This mapping is well defined because it does not depend on the choice of the representation $v$ of the class $U$ by virtue of $\theta$-compatibility of $P_{B_\theta}$. Indeed, since $uP_{B_\theta}v$, then $\theta(u)P_{B_\theta}\theta(v)$ and hence $\theta(v) \in \psi(U)$. Therefore if $V$ is the class of $P_{B_\theta}$ containing $v$, then $\psi(U) = \psi(V)$. It is immediate that $\psi$ is an antimorphism since $\theta$ is an antimorphism. To show that $\psi$ is an involution, for all $U \in Syn(B_\theta)$, $\psi(\psi(U)) = U$. Note that $\psi(U) = [\theta(u)]$ for all $u \in U$. Thus $\psi(\psi(U)) = [\theta(\theta(u))] = [u] = U$ since $\theta$ is an involution. Thus $\psi$ is an antimorphic involution. The last part follows from the fact that $B_\theta$ is $\theta$-stale.
3. Let $D_1 = \{[x] \in Syn(B_\theta) : x \in B_\theta\}$ and let $D_2 = Syn(B_\theta) \setminus D_1 = \{[x] \in Syn(B_\theta) : x \in \overline{B_\theta}\}$. Let $[x] \in D_1$ which implies that $x \in B_\theta$ and thus $x = arb$ for some $a, b \in \Sigma$ and $r \in \Sigma^*$ with $\theta(a) = b$. Thus from Corollary 4, we have $\psi([x]) = \psi([arb]) = \psi([ab]) = [\theta(ab)] = [\theta(b)\theta(a)] = [ab] = [x]$. Thus for all $[x] \in D_1$, $\psi([x]) = [x]$. Now we show that $D_1$ is disjunctive. Suppose there exists $[x], [y] \in Syn(B_\theta)$ such that $C_{D_1}([x]) = C_{D_1}([y])$. Then $[\alpha][x][\beta] \in D_1$ iff $[\alpha][y][\beta] \in D_1$ for $[\alpha], [\beta] \in Syn(B_\theta)$ which implies that $[\alpha x\beta] \in D_1$ iff $[\alpha y\beta] \in D_1$. Thus for all $\alpha, \beta \in \Sigma^*$, $\alpha x\beta \in B_\theta$ iff $\alpha y\beta \in B_\theta$ which implies $C_{B_\theta}(x) = C_{B_\theta}(y)$ and hence $x, y \in [x] = [y]$. Hence $D_1$ is disjunctive. Since $P_{D_1} = P_{\overline{D_1}} = P_{D_2}$, $D_2$ is also disjunctive.

The next proposition is a converse of the Proposition 5.

**Proposition 6** *Let $M$ be a monoid with identity $e$ and satisfy the following properties:*

1. *$M$ is finite.*
2. *$M$ has no zero.*
3. *Every element of $M$ is an idempotent element.*
4. *There exists an antimorphic involution $\psi$ such that $M$ is stable under $\psi$.*
5. *$M$ has two non empty disjunctive subsets $D_1$ and $D_2$ such that $D_1 = \{x \in M \setminus \{e\} : \psi(x) = x\}$ and $D_2 = M \setminus D_1$ and for all $x \in D_1$ there exists $p, q, r \in D_2$ such that $x = pq$ and either $\psi(p) = rq$ or $\psi(q) = pr$.*

*Then there exists a free monoid $\Sigma^*$ over a finite alphabet $\Sigma$, an antimorphic involution $\theta$ and a language $B_\theta$ in $\Sigma^*$ such that,*
*(i) $B_\theta$ is the set of all $\theta$-bordered words over $\Sigma$*
*(ii) The syntactic monoid $Syn(B_\theta) = \Sigma^*/P_{B_\theta}$ is isomorphic to $M$.*

*Proof.* If $M = \{x_1, x_2, ..., x_n\}$, then take the elements of $M$ as the letters of an alphabet $\Sigma = \{x_1, x_2, ..., x_n\}$ and let $\Sigma^*$ be the free monoid generated by $\Sigma^*$. Let $\phi$ be the mapping of $\Sigma^*$ onto $M$ defined in the following way. If $u \in \Sigma$, then $\phi(u) = \psi(u) \in M$.

If $u = u_1u_2...u_k \in \Sigma^+$ with $u_i \in \Sigma$, then $\phi(u) = \psi(u) = \psi(u_k)...\psi(u_1)$. If $u = \lambda$, then $\phi(u) = e$, the identity of $M$. It is clear that $\phi$ is an antimorphism on $\Sigma^*$ onto $M$. The relation $\rho$ defined as $u\rho v$, $u, v \in \Sigma^*$ iff $\phi(u) = \phi(v)$ is a congruence of $\Sigma^*$ and the quotient monoid $\Sigma^*/\rho$ is isomorphic to $M$.
Let $B_\theta = \{x \in \Sigma^+ : \phi(x) = x\}$ and let $P_{B_\theta}$ be the syntactic congruence of $B_\theta$. We need to show that $P_{B_\theta} = \rho$. We first show that $\rho \subseteq P_{B_\theta}$. Let $u\rho v$ then $\phi(u) = \phi(v)$. We need to show that $uP_{B_\theta}v$. Let $\alpha u\beta \in B_\theta$ which implies that $\phi(\alpha u\beta) = \alpha u\beta = \phi(\beta)\phi(u)\phi(\alpha) = \phi(\beta)\phi(v)\phi(\alpha)$ since $u\rho v$. Thus $\phi(\alpha u\beta) = \phi(\alpha v\beta) = \phi(\alpha u\beta)$ which implies $\alpha v\beta = \alpha u\beta$, since $\phi$ is an involution, it is bijective. Thus $\phi(\alpha v\beta) = \alpha v\beta$ which implies that $\alpha v\beta \in B_\theta$. Similarly we can show that $\alpha v\beta \in B_\theta$ and hence $\alpha u\beta \in B_\theta$. Thus $uP_{B_\theta}v$.
Conversely, we need to show that $P_{B_\theta} \subseteq \rho$. Let $uP_{B_\theta}v$. If $u$ is not equivalent to $v$ modulo $\rho$ then $\phi(u) \neq \phi(v)$. $M$ has a disjunctive $D_1$. Then syntactic congruence $P_{D_1}$ is the equality relation and we have $C_{D_1}(\phi(u)) \neq C_{D_1}(\phi(v))$. This implies the existence if $\alpha, \beta \in M$ such that $\alpha\phi(u)\beta \in D_1$ and $\alpha\phi(v)\beta \notin D_1$ or $\alpha\phi(u)\beta \notin D_1$ and $\alpha\phi(v)\beta \in D_1$. Suppose that we have the first case, $\alpha\phi(u)\beta \in D_1$ and $\alpha\phi(v)\beta \notin D_1$, and since $\phi$ is bijective there exists $r, s \in \Sigma^*$ such that $\alpha = \phi(r)$, and $\beta = \phi(s)$. Thus $\alpha\phi(x)\beta = \phi(r)\phi(u)\phi(s) = \phi(sur) \in D_1$ and $\phi(svr) \notin D_1$, i.e., $\phi(sur) = sur$ and $\phi(svr) \neq svr$ which implies that $sur \in B_\theta$ and $svr \notin B_\theta$ a contradiction since $C_{B_\theta}(u) = C_{B_\theta}(v)$. Hence it follows that $P_{B_\theta} \subseteq \rho$.
We define the requested antimorphism $\theta$ of $\Sigma^*$ by taking the corresponding permutation of the alphabet $\Sigma$ and extending it to $\Sigma^*$ in the usual way. If $u \in \Sigma^+$, $u = x_1x_2...x_n$ for $x_1, x_2, ..., x_n \in \Sigma$, then $\theta(u) = \theta(x_1x_2...x_n) = \theta(x_n)...\theta(x_1)$ and $\theta(\lambda) = \lambda$. It is immediate that $\theta$ is bijective antimorphism. Let us show now that conditions $(i)$ and $(ii)$ are satisfied.
For $(i)$, let $u \in B_\theta$ and suppose that $u$ is $\theta$-unbordered. If $u \in B_\theta$ then $u = u_1u_2...u_k$ for some $u_i \in \Sigma$. Then if a word $u$ is Watson-Crick bordered and its WK borders overlap, the word $u$ may stick to another copy of itself as shown above. $\phi(u) = \phi(u_k)...\phi(u_1) = u_1...u_k$ which implies $\phi(u_k)\phi(u_1) = u_1u_k$ by Corollary 4. Thus $u_1 = u_k$. Hence $\phi(u) = \phi(u_1u_k) = \phi(u_1u_1) = \phi(u_1) = u = u_1u_1 = u_1$ since for all $f \in M$, $f$ is an idempotent. Thus for all $u \in B_\theta$, $u = u_1$ for some $u_1 \in D_1$. Hence there exists $p, q, r \in D_2$ such that $u = pq$ with $\psi(p) = rq$ or $\psi(q) = pr$. Thus $u = pq$ implies $\psi(u) = \psi(q)\psi(p) = pr\psi(p)$ or $\psi(u) = \psi(q)rq$ which implies that $u$ is $\theta$-bordered. Suppose there exists a $u \in \Sigma^*$ such that $u$ is $\theta$-bordered and $u \notin B_\theta$, then $u = axb$ with $\theta(a) = b$ and $a, b \in \Sigma$. Thus $\psi(u) = \psi(axb) = \psi(ax\theta(a)) = \psi(\theta(a))\psi(x)\psi(a) = \psi(\theta(a))\psi(a) = \psi(b)\psi(a) = ab = axb = u$. Thus $\psi(u) = u$ implies that $\phi(u) = u$ and $u \in B_\theta$.
Condition $(ii)$ follows by construction. □

## 4 Green's relations for the set of all Watson-Crick bordered words

We recall here the definition of Green's relations and some well known facts about some of the relations. For extensive treatments of Green's relations and the related varieties of finite monoids, we refer the reader to [4, 12, 17]. In [10], it was shown that Green's relations are trivial for the language of all hairpin-free words. In contrast, this is not the case for the language of all involution-bordered words. Namely, in this section we show that $S = Syn(B_\theta) \setminus \{1\}$ is $\mathcal{H}$-trivial and $S$ is not $\mathcal{K}$-trivial for all $\mathcal{K} \in \{\mathcal{D}, \mathcal{R}, \mathcal{L}, \mathcal{J}\}$.

**Definition 3.** (Green's Relations:) *Let $S$ be a semigroup. We define on $S$ four equivalence relations $\mathcal{R}$, $\mathcal{L}$, $\mathcal{H}$ and $\mathcal{J}$ called Green's relations:*

$$a\mathcal{R}b \Leftrightarrow aS^1 = bS^1$$
$$a\mathcal{L}b \Leftrightarrow S^1a = S^1b$$
$$a\mathcal{J}b \Leftrightarrow S^1aS^1 = S^1bS^1$$
$$a\mathcal{H}b \Leftrightarrow a\mathcal{R}b \text{ and } a\mathcal{L}b$$

Note that the relations $\mathcal{R}$ and $\mathcal{L}$ commute, i.e., $\mathcal{RL} = \mathcal{LR}$ and $\mathcal{D} = \mathcal{RL}$. In a finite semigroup $\mathcal{D} = \mathcal{J}$. A semigroup $S$ is $\mathcal{K}$-trivial iff $e\mathcal{K}f$ implies $e = f$ for $\mathcal{K} \in \{\mathcal{D}, \mathcal{R}, \mathcal{L}, \mathcal{J}, \mathcal{H}\}$. A semigroup $S$ is aperiodic if for all $x \in S$ there exists $n$ such that $x^n = x^{n+1}$. Note that $S = Syn(B_\theta) \setminus \{1\}$ is aperiodic since all elements of $S$ are idempotents.

We use the following propositions from [17] to show that $S = Syn(B_\theta) \setminus \{1\}$ is $\mathcal{H}$-trivial and the $\mathcal{D}$ class of $S$ is equal to $S$.

**Proposition 7** *[17] Let $S$ be a semigroup and let $g$ and $f$ be idempotents of $S$. Then $g\mathcal{D}f$ if and only if $g$ and $f$ are conjugates, i.e., there exists $u, v \in S$ such that $g = uv$ and $f = vu$.*

**Proposition 8** *[17] Let $S$ be a finite semigroup. The following conditions are equivalent.*
*1. $S$ is aperiodic (for every $x \in S$ there exists $n$ such that $x^n = x^{n+1}$).*
*2. There exists $m > 0$ such that for every $x \in S$, $x^m = x^{m+1}$.*
*3. $S$ is $\mathcal{H}$-trivial.*

**Proposition 9** *The $\mathcal{D}$ class and $\mathcal{J}$ class of $S$ is equal to $S$.*

*Proof.* Follows from the fact that $S$ is simple and finite.

**Proposition 10** *$S = Syn(B_\theta) \setminus \{1\}$ is $\mathcal{H}$-trivial.*

*Proof.* Since $S$ is aperiodic, by Proposition 8, $S$ is $\mathcal{H}$-trivial. □

**Proposition 11** *Let $\Sigma = \{a_1, a_2, ..., a_n\}$ and let $[ab] \in S = Syn(B_\theta) \setminus \{1\}$ for some $a, b \in \Sigma$. Then the $\mathcal{R}$ class of $[ab]$ is $\{[aa_i] : a_i \in \Sigma\}$. and $\mathcal{L}$ class of $[ab]$ is $\{[a_ib] : a_i \in \Sigma\}$.*

*Proof.* Let $e\mathcal{R}f$ where $e = [ab]$ for some $a, b \in \Sigma$. $e = [ab]$ is the set of all words that begin with $a$ and end with $b$. Then for all $f \in [ab]S^1$, $f$ is the set of all words that begin with $a$. Thus $\mathcal{R}$ class of $[ab]$ is $\{[aa_i] : a_i \in \Sigma\}$. Similarly we can show that the $\mathcal{L}$ class of $[ab]$ is $\{[a_ib] : a_i \in \Sigma\}$. □

**Corollary 7** *For all $e, f \in S$, $\mathcal{R}_e \cap \mathcal{L}_f = \{ef\}$.*

*Proof.* For some $e, f \in S$, $e = [ae_1]$ and $f = [f_1b]$ for some $a, b \in \Sigma$ and $e_1, f_1 \in \Sigma^*$. Note that $ef = [ae_1].[f_1b] = [ae_1f_1b] = [ab]$. Then from Proposition 11, $\mathcal{R}_{[ae_1]} = \{[aa_i] : a_i \in \Sigma\}$ and $\mathcal{L}_{[f_1b]} = \{[a_ib] : a_i \in \Sigma\}$. Thus $\mathcal{R}_e \cap \mathcal{L}_f = \{[ab]\} = \{ef\}$.

*Example 1.* Let $\Delta = \{A, C, G, T\}$ and let $\theta$ be an antimorphic involution that maps $A \mapsto T$ and $C \mapsto G$. Then $B_\theta = \{aub : a, b \in \Delta, \theta(a) = b, u \in \Delta^*\}$ is the set of all $\theta$-bordered words over $\Delta^*$. Then $Syn(B_\theta) = \{[1], [A], [C], [G], [T], [AC], [AG], [AT], [CA], [CG], [CT], [GA], [GC], [GT], [TA], [TG], [TC]\}$. Note that for all $a, b \in \Delta$, $[ab]$ is the set of all words that begin with $a$ and end with $b$ and $[a]$ represents the class that contains all words that begin and end with $a$. We now compute both the $\mathcal{R}$ and $\mathcal{L}$ class for elements of $Syn(B_\theta)$.

- $\mathcal{L}_{[A]} = \{\ [A], [CA], [GA], [TA]\}$
- $\mathcal{L}_{[C]} = \{\ [C], [AC], [GC], [TC]\}$
- $\mathcal{L}_{[G]} = \{\ [G], [AG], [CG], [TG]\}$
- $\mathcal{L}_{[T]} = \{\ [T], [CT], [GT], [AT]\}$
- $\mathcal{R}_{[A]} = \{\ [A], [AC], [AG], [AT]\}$
- $\mathcal{R}_{[C]} = \{\ [C], [CA], [CG], [CT]\}$
- $\mathcal{R}_{[G]} = \{\ [G], [GA], [GC], [GT]\}$
- $\mathcal{R}_{[T]} = \{\ [T], [TA], [TG], [TC]\}$

Also note that since $\mathcal{H} = \mathcal{R} \cap \mathcal{L}$ for all $e \in Syn(B_\theta)$, $\mathcal{H}_e = \{e\}$

## 5 Conclusion

The DNA secondary structure called "hairpin" has been a topic of constant interest in experimental as well as theoretical biomolecular computing, as it is usually undesirable in DNA-based computing experiments. This paper investigates a mathematical formalization of a particular case of hairpins, the Watson-Crick bordered words, whereby the "sticky borders" that cause a DNA single strand to form a hairpin are situated at the extremities of the strand. Cases where these "sticky borders" are situated in the interior of the strand have been addressed, e.g., in [9], [10]. The main results of this paper are algebraic properties of Watson-Crick bordered and unbordered words, and a complete characterization of the syntactic monoid of the language consisting of all Watson-Crick bordered words over a given alphabet.

Directions for future work are two-fold. On one hand we intend to investigate other generalizations of classical notions in combinatorics of words motivated by DNA computing, such as Watson-Crick conjugate words and Watson-Crick commutative words. On the other hand, we intend to formalize other DNA secondary structures such as DNA pseudo-knots and study their properties.

**Acknowledgment** Research supported by NSERC and Canada Research Chair grants for Lila Kari.

## References

1. J.Berstel and D.Perrin, *Theory of Codes,* Academic Press, Inc. Orlando Florida, (1985).
2. M.Domaratzki. *Hairpin structures defined by DNA trajectories*, Proc. of DNA Computing 12, C.Mao, T.Yokomori, Editors, LNCS 4287 (2006), 182-194.
3. M.Garzon, V.Phan, S.Roy and A.Neel. *In search of optimal codes for DNA computing*, Proc. of DNA Computing 12, C.Mao, T.Yokomori, Editors, LNCS 4287 (2006), 143-156.
4. J.M.Howie, *Fundamentals of Semigroup Theory*, Oxford Science Publications, (1995).
5. N.Jonoska, K.Mahalingam and J.Chen, *Involution codes: with application to DNA coded languages*, Natural Computing, Vol 4-2 (2005), 141-162.
6. H.Jürgensen, *Syntactic monoid of codes*, Acta Cybernetica 14 (1999), 117-133.
7. H.Jürgensen and S.Konstantinidis, *Codes*, Handbook of Formal Languages, Vol 1, Chapter 8, G.Rozenberg, A.Salomaa, Editors, (1997), 511-608.
8. L.Kari, S.Konstantinidis, E.Losseva and G.Wozniak, *Sticky-free and overhang-free DNA languages*, Acta Informatica 40 (2003), 119-157.
9. L.Kari, S.Konstantinidis, E.Losseva, P.Sosik and G.Thierrin, *Hairpin structures in DNA words*, Proceedings of DNA Computing 11, A.Carbone, N.Pierce, Editors, LNCS 3892 (2005), 158-170.
10. L.Kari, K.Mahalingam and G.Thierrin, *The syntactic monoid of hairpin-free languages*, Accepted, Acta Informatica (2007).
    Available online: http://www.springerlink.com/content/2r264425831k6283/
11. L.Kari and K.Mahalingam, *Involution bordered words*, Accepted, IJFCS, (2007). Available online: http://www.csd.uwo.ca/~ lila/invbor.pdf

12. G.Lallement, *Semigroups and Combinatorial Dynamics*, Wiley/Interscience, New York (1995).
13. A.De Luca and A.Restivo, *A characterization of strictly locally testable languages and its application to subsemigroups of a free semigroup*, Information and Control 44 (1980), 300-319.
14. M.Petrich and C.M.Reis, *The syntactic monoid of the semigroup generated by a comma-free code*, Proceedings of the Royal Society of Edinburgh, 125A (1995), 165-179.
15. M.Petrich, C.M. Reis and G.Thierrin, *The syntactic monoid of the semigroup generated by a maximal prefix code*, Proceedings of the American Mathematical Society, 124-3 March (1996), 655-663.
16. M.Petrich and G.Thierrin, *The syntactic monoid of an infix code*, Proceedings of the American Mathematical Society 109-4 (1990), 865-873.
17. J.E.Pin, *Varieties of Formal Languages*, Plenum Press (1986).
18. H.J.Shyr, *Free Monoids and Languages*, Hon Min Book Company (2001).
19. G.Thierrin, *The syntactic monoid of a hypercode*, Semigroup Forum 6 (1973), 227-231.
20. S.S.Yu, *d-Minimal Languages*, Discrete Applied Mathematics 89 (1998), 243-262.