# Hairpin Structures Defined by DNA Trajectories[*]

Michael Domaratzki
Department of Computer Science,
University of Manitoba,
Winipeg, MB R3T 2T2 Canada
mdomarat@cs.umanitoba.ca

### Abstract

We examine scattered hairpins, which are structures formed when a single strand of nucleotides folds into a partially hybridized stem and a loop. To specify different classes of hairpins, we use the concept of DNA trajectories, which allows precise descriptions of valid bonding patterns on the stem of the hairpin. DNA trajectories have previously been used to describe bonding between separate strands.

We are interested in the mathematical properties of scattered hairpins described by DNA trajectories. We examine the complexity of the set of hairpin-free words described by a set of DNA trajectories. In particular, we consider the closure properties of language classes under sets of DNA trajectories of differing complexity. We address decidability of recognition problems for hairpin structures.

## 1 Introduction

A hairpin in a single strand of nucleotides is a structure formed by the bonding of two complementary regions, which form the *stem*, joined on one end by an intermediate, unbonded region. Together, the stem and the unbonded region (the *loop*) are known as a hairpin. We illustrate this concept in Figure 1.

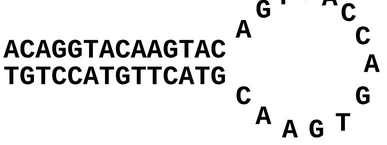

Figure 1: A hairpin in a strand of nucleotides.

[*]An earlier version of this paper without proofs appeared at DNA 12 [5]. Research conducted at the Jodrey School of Computer Science, Acadia University, and supported in part by a grant from NSERC.

As research into DNA computing applications and nanotechnology continues, the formal study of hairpins gains increasing significance. Kari *et al.* [12, 14, 15] survey the use of hairpins in various contexts. We also note the use of hairpins for visual contrast in evaluating successful nanotechnological constructions, as described in, e.g., the work of Rothemund *et al.* [22]. In some of these applications, hairpins are desirable, while in other applications, they are problematic and are to be avoided in sets of DNA strands. Further, hairpins serve as the basis for more complicated secondary structures such as pseudoknots.

Recently, Kari *et al.* [12, 14, 15] have studied hairpins using the tools of theoretical computer science. In particular, a single strand of nucleotides is viewed as a word over the alphabet $\Delta = \{A,C,G,T\}$. In this framework, a hairpin in a word $z$ is a decomposition $z = uvwxy$ where $v$ and $x$ are complementary to each other, and form the stem of the hairpin. We characterize the complementarity of $v$ and $x$ using an antimorphism $\theta$ (for definitions, see Section 2). Among other results, Kari *et al.* characterize the complexity and decidability results for hairpin sets [14]. Further, Kari *et al.* [12] have also studied *scattered* hairpins, which represent hairpins in which the stem is not completely hybridized, i.e., where an arbitrary number of unbonded regions occur within the stem.

In this paper, we examine refinements of hairpins and scattered hairpins by incorporating a parameter—a set of *DNA trajectories*—to add increased capability in describing the set of hairpins which are of potential interest. The use of DNA trajectories has recently been employed to model bonding regions in separate strands, called *bond-free properties* [13].

The introduction of DNA trajectories in this paper as a refinement of the results of Kari *et al.* has several advantages. One main benefit of DNA trajectories is that they enable constraints to be expressed as a formal language, rather than graphically or otherwise. DNA trajectories also allow more precise specifications of the form of DNA hairpins we are interested in than previous work, which allows the tools developed in this paper to be applied to more complex DNA computing models. Further, DNA trajectories are capable of adapting to minor structural changes: modifications such as enforcing a minimum length of a bond are easily introduced in DNA trajectories, instead of as a separate specification (the technique adopted by Kari *et al.*). Beyond the use of DNA trajectories to aid in modelling situations of practical importance, we follow the work of Kari *et al.* and examine not only sets which allow the presence of certain hairpin formations, but hairpin-free sets, where we stipulate that hairpins from a given specification set cannot occur.

In our study of DNA trajectories and hairpins, we focus on closure properties, decidability and relations to problems from combinatorics on words. With respect to closure properties, we find that the addition of DNA trajectories gives a more complex situation than the case of hairpins and scattered hairpins studied by Kari *et al.*, and many results have been obtained. In particular, we find that by allowing a set of DNA trajectories, we cannot guarantee that the set of all hairpins will still form a regular language, and several conditions are investigated which yield interesting theoretical insights. Decidability problems are also more interesting, due to the fact that regularity of a set of DNA trajectories does not imply the regularity of the associated set of hairpins or the set of hairpin-free DNA words.

## 2 Definitions

For additional background in formal languages and automata theory, please see Rozenberg and Salomaa [23]. For an introduction to DNA computing, see Păun *et al.* [18]. Let $\Sigma$ be a finite set of symbols, called *letters*; we call $\Sigma$ an alphabet. Then $\Sigma^*$ is the set of all finite sequences of letters from $\Sigma$, which are called *words.* The empty word $\varepsilon$ is the empty sequence of letters. The *length* of a word $w = w_1 w_2 \cdots w_n \in \Sigma^*$, where $w_i \in \Sigma$, is $n$, and is denoted $|w|$. Given a word $w \in \Sigma^*$ and $a \in \Sigma$, $|w|_a$ is the number of occurrences of $a$ in $w$. Given two words $x = x_1 x_2 \cdots x_n$ and $y = y_1 y_2 \cdots y_m$ where $x_i, y_j \in \Sigma$ for $1 \leq i \leq n$ and $1 \leq j \leq m$, the *concatenation* of $x$ and $y$ is denoted by $xy$ and is given by $xy = x_1 x_2 \cdots x_n y_1 y_2 \cdots y_m$.

A *language* $L$ is any subset of $\Sigma^*$. Given languages $L_1, L_2 \subseteq \Sigma^*$, their concatenation is defined by $L_1 L_2 = \{xy \ : \ x \in L_1, y \in L_2\}$. We define powers of languages by $L^0 = \{\varepsilon\}$ and $L^i = L^{i-1}L$ for all languages $L$ and all $i \geq 1$. By $L^*$ we mean $\cup_{i \geq 0} L^i$.

We use the notation $\prod_{i=1}^{n} L_i$ to denote $L_1 L_2 \cdots L_n$, and the notation $L^{\geq k}$ to denote $L^k L^*$. The reversal of a word $w = x_1 x_2 \cdots x_n$ ($x_i \in \Sigma$), denoted $w^R$, is defined by $w^R = x_n \cdots x_2 x_1$. By extension, $L^R = \{x^R \ : \ x \in L\}$.

Let $\Sigma, \Delta$ be alphabets and $h : \Sigma \to \Delta^*$ be any function. Then $h$ can be extended to a morphism $h : \Sigma^* \to \Delta^*$ via the condition that $h(uv) = h(u)h(v)$ for all $u, v \in \Sigma^*$. Similarly, $h$ can be extended to an antimorphism via the condition that condition that $h(uv) = h(v)h(u)$ for all $u, v \in \Sigma^*$. An involution $\theta$ is any function $\theta : \Sigma \to \Sigma$ such that $\theta^2$ is the identity function on $\Sigma$. Let $\mu$ denote the mirror involution (i.e., the identity function extended to an antimorphism). Let $\iota$ denote the identity morphism.

Given alphabets $\Sigma, \Delta$, a substitution is any function $h : \Sigma \to 2^{\Delta^*}$. It is extended to $h : \Sigma^* \to 2^{\Delta^*}$ by the condition that $h(uv) = h(u)h(v)$ for all $u, v \in \Sigma^*$. A substitution is finite if $h(a)$ is a finite language over $\Delta$ for all $a \in \Sigma$.

A *deterministic finite automaton* (DFA) is a five-tuple $M = (Q, \Sigma, \delta, q_0, F)$ where $Q$ is the finite set of states, $\Sigma$ is the alphabet, $\delta : Q \times \Sigma \to Q$ is the (partial) transition function, $q_0 \in Q$ is the start state, and $F \subseteq Q$ is the set of final states. We extend $\delta$ to $Q \times \Sigma^*$ in the usual way. A word $w \in \Sigma^*$ is accepted by $M$ if $\delta(q_0, w) \in F$. The *language accepted* by $M$, denoted $L(M)$, is the set of all words accepted by $M$. A language is called *regular* if it is accepted by some DFA.

A *context-free grammar* (CFG) is a four-tuple $G = (V, \Sigma, P, S)$, where $V$ is a finite set of non-terminals, $\Sigma$ is a finite alphabet, $P \subseteq V \times (V \cup \Sigma)^*$ is a finite set of productions and $S \in V$ is the start non-terminal. If $(\alpha, \beta) \in P$, we usually denote this by $\alpha \to \beta$. A CFG is *linear* (an LCFG) if $P \subseteq V \times (\Sigma^*(V \cup \{\varepsilon\})\Sigma^*)$. A CFG is *left-linear* if $P \subseteq V \times (\Sigma^*(V \cup \{\varepsilon\}))$. It is known that we can assume without loss of generality that the productions in a left-linear grammar $G$ are of form $P \subseteq V \times (\Sigma(V \cup \{\varepsilon\}))$ if $\varepsilon \notin L(G)$.

If $G = (V, \Sigma, P, S)$ is a CFG, then given two words $\alpha, \beta \in (V \cup \Sigma)^*$, we denote $\alpha \Rightarrow_G \beta$ if $\alpha = \alpha_1 \alpha_2 \alpha_3$, $\beta = \alpha_1 \beta_2 \alpha_3$ for $\alpha_1, \alpha_2, \alpha_3, \beta_2 \in (V \cup \Sigma)^*$ and $\alpha_2 \to \beta_2 \in P$. Let $\Rightarrow_G^*$ denote the reflexive, transitive closure of $\Rightarrow_G$. Then the language generated by a grammar $G = (V, \Sigma, P, S)$ is given by $L(G) = \{x \in \Sigma^* \ : \ S \Rightarrow_G^* x\}$. If a language is generated by a CFG (resp., LCFG), then it is a context-free language (CFL) (resp., linear context-free language (LCFL)). The class of languages accepted by left-linear grammars are known to be exactly the regular languages.

## 2.1 Trajectory-based Operations

The shuffle on trajectories operation is a method for specifying the ways in which two input words may be interleaved to form a result. Each trajectory $t \in \{0,1\}^*$ with $|t|_0 = n$ and $|t|_1 = m$ (i.e., with $n$ occurrences of 0 and $m$ occurrences of 1) specifies one particular way in which we can shuffle two words of length $n$ (as the left input word) and $m$ (as the right input word). The word of length $n+m$ resulting from the shuffle along $t$ will have a letter from the left input word in position $i$ if the $i$-th symbol of $t$ is 0, and a letter from the right input word in position $i$ if the $i$-th symbol of $t$ is 1.

The formal definition is given as follows [17]:

**Definition 2.1.** Let $x$ and $y$ be words over an alphabet $\Sigma$ and $t$, the *trajectory*, be a word over $\{0,1\}$. The shuffle of $x$ and $y$ on trajectory $t$ is denoted by $x \sqcup\!\sqcup_t y$. If $t = \prod_{i=1}^{n} 0^{j_i} 1^{k_i}$ for some $n \geq 0$ and $j_i, k_i \geq 0$ for all $1 \leq i \leq n$, then

$$x \sqcup\!\sqcup_t y = \{\prod_{i=1}^{n} x_i y_i \ : x = \prod_{i=1}^{n} x_i, y = \prod_{i=1}^{n} y_i, \text{ with } |x_i| = j_i, |y_i| = k_i \text{ for all } 1 \leq i \leq n\}$$

if $|x| = |t|_0$ and $|y| = |t|_1$, and $x \sqcup\!\sqcup_t y = \emptyset$ if $|x| \neq |t|_0$ or $|y| \neq |t|_1$. We extend the operation of shuffle on trajectories to sets of trajectories $T \subseteq \{0,1\}^*$ as follows:

$$x \sqcup\!\sqcup_T y = \bigcup_{t \in T} x \sqcup\!\sqcup_t y.$$

Further, if $L_1, L_2 \subseteq \Sigma^*$ are languages, then

$$L_1 \sqcup\!\sqcup_T L_2 = \bigcup_{\substack{x \in L_1 \\ y \in L_2}} x \sqcup\!\sqcup_T y.$$

As an example, note that if $T = 0^*1^*$, then $\sqcup\!\sqcup_T$ is the concatenation operation: $L_1 \sqcup\!\sqcup_T L_2 = L_1 L_2$. If $T = 0^*1^*0^*$, then $\sqcup\!\sqcup_T$ is the insertion operation $\leftarrow$, defined by $L_1 \leftarrow L_2 = \{x_1 y x_2 \ : \ x_1 x_2 \in L_1, y \in L_2\}$.

We will also require the notion of the natural binary relation defined by shuffle on trajectories [3]. For $T \subseteq \{0,1\}^*$, define $\omega_T$ as follows: for all $x, y \in \Sigma^*$, $x \, \omega_T \, y \iff y \in x \sqcup\!\sqcup_T \Sigma^*$.

For example, if $T = 0^*1^*$, then $\omega_T$ is the prefix order, defined by $x \, \omega_T \, y$ if and only if $y \in x\Sigma^*$. If $T = \{0,1\}^*$, then $x \, \omega_T \, y$ is the embedding order, defined by $x \, \omega_T \, y$ if and only if $y \in x \sqcup\!\sqcup \Sigma^*$ (i.e., $x$ can be obtained from $y$ by deleting zero or more letters). We denote the embedding order by $\leq_e$; note that if $x \leq_e y$ then $x$ is a *scattered subword* of $y$.

## 2.2 DNA Trajectories and Hairpins

We now consider DNA trajectories, defined by Kari *et al.* [13]. A DNA trajectory is a word over the alphabet $V_D = \left\{ \binom{b}{b}, \binom{f}{f}, \binom{f}{\varepsilon}, \binom{\varepsilon}{f} \right\}$. The original use of a set of DNA trajectories was to define bonding between two separate single strands of DNA. The occurrence of $\binom{b}{b}$ implies a bond at a certain position, while $\binom{f}{f}$ (resp., $\binom{f}{\varepsilon}$, $\binom{\varepsilon}{f}$) denotes two bases which are free (resp., an extra unbonded nucleotide on the top strand, an extra unbonded nucleotide on the bottom strand). DNA trajectories are used to define so-called *bond-free properties* in DNA code word design [13], and we adopt them here for modelling the bonding of hairpins.

For hairpins, we can view words over $V_D^*$ as designating where bonds can occur and cannot occur when viewing the strands with the loop at the right end. For instance, the DNA trajectory $t = \binom{f}{\varepsilon}\binom{f}{f}^2\binom{b}{b}^3\binom{f}{f}^3\binom{f}{\varepsilon}$ represents the bonding depicted in Figure 2. Note that the pairs $x_4$ and $x_{16}$, $x_5$ and $x_{15}$, as well as $x_6$ and $x_{14}$ must be bonded together (this assumes an antimorphic bonding pattern–see Definition 2.2 below).

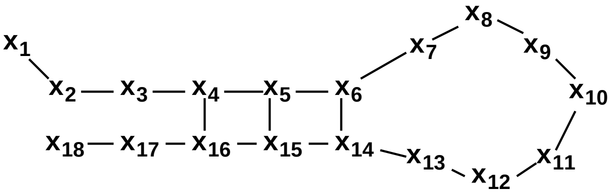

Figure 2: A DNA bond specified by $t$. The letters $x_i$ represent arbitrary letters from the alphabet.

Let $\varphi_u, \varphi_d : V_D^* \to \{0,1\}^*$ be morphisms defined by

$$\begin{array}{rclrclrcl}
\varphi_u(\binom{b}{b}) & = & 0, & \varphi_u(\binom{f}{y}) & = & 1, & \text{for } y \in \{f,\varepsilon\}, \quad \varphi_u(\binom{\varepsilon}{f}) & = & \varepsilon, \\
\varphi_d(\binom{b}{b}) & = & 0, & \varphi_d(\binom{y}{f}) & = & 1, & \text{for } y \in \{f,\varepsilon\}, \quad \varphi_d(\binom{f}{\varepsilon}) & = & \varepsilon.
\end{array}$$

We now give our main definition.

**Definition 2.2.** Let $\Sigma$ be an alphabet, $\theta : \Sigma \to \Sigma$ be an arbitrary involution, extended to a morphism or antimorphism, and $S \subseteq V_D^*$. Then a word $w$ is said to be $S$-$\theta$-hairpin-free, or simply $shp_\Sigma(S,\theta)$-free, if the following condition holds

$$\forall u,v,x,\in \Sigma^*, s \in S, (w = uv, x\,\omega_{\varphi_u(s)}\,u, \text{ and } \theta(x)\,\omega_{\varphi_d(s)^R}\,v) \Rightarrow x = \varepsilon. \tag{1}$$

That is, $w$ is $S$-$\theta$-hairpin free if we can write $w$ as $w = uv$ and there exists a word $x$—which represents the portions of $u$ and $v$ which are bonded together—such that

(1) $x$ appears in $u$ according to the bonding prescribed by $\varphi_u(s)$ and

(2) $\theta(x)$ appears in $v$ according to the bonding prescribed by $\varphi_d(s)^R$.

then $x = \varepsilon$. Note that $\varphi_d(s)$ is reversed since $v$ runs backwards from the right-to-left in our hairpin.

**Example 2.3.** Let $\Sigma = \{a,b,c\}$, and $t = \binom{f}{\varepsilon}\binom{f}{f}^2\binom{b}{b}^3\binom{f}{f}^3\binom{f}{\varepsilon}$ be the DNA trajectory from Figure 2. Note that $\varphi_u(t) = 1110001111$ and $\varphi_d(t)^R = 11100011$.

In this case, $w = a^3baca^7cabb^2$ is not $shp_\Sigma(\{t\},\mu)$-free (recall that $\mu$ is the identity antimorphism) since the conditions of (1) are violated with $u = a^3baca^4$, $v = a^3cabb^2$ and $x = bac$. However, we can verify that $w = a^3baca^7baac^2$ is $shp_\Sigma(\{t\},\mu)$-free.

**Definition 2.4.** We say that a language $L$ is $shp_{\Sigma}(S,\theta)$-free if $w$ is $shp_{\Sigma}(S,\theta)$-free for all $w \in L$.

Let $shpf_{\Sigma}(S,\theta)$ denote the set of $shp_{\Sigma}(S,\theta)$-free words. Let $shp_{\Sigma}(S,\theta) = \Sigma^* - shpf_{\Sigma}(S,\theta)$. Clearly, $L$ is $shp_{\Sigma}(S,\theta)$-free if and only if $L \subseteq shpf_{\Sigma}(S,\theta)$.

The definition of $shp_{\Sigma}(S,\theta)$-freeness is an extension of the notions of hairpin-freeness and scattered-hairpin-freeness, investigated by Kari *et al.* [12, 14].

Note that in the above definition $\theta$ can be an arbitrary involution, extended to either a morphism or antimorphism. This is similar to the work on bond-free properties [13] and hairpin-freeness [12, 14]. In practice, an antimorphic involution yields results applicable to hairpin and scattered-hairpin structures, while morphic involutions yield structures where the scattered stem is bonded in a parallel, rather than an anti-parallel, orientation. Of course, the antimorphic involution $\tau$ over the alphabet $\Delta = \{A,C,G,T\}$ defined by $\tau(A) = T, \tau(T) = A, \tau(C) = G$ and $\tau(G) = C$ is of particular interest in practice. This involution is called the Watson-Crick involution. In biological settings, only anti-parallel orientations arise, so the case where the involution $\theta$ is extended to an antimorphism models this situation; the case of morphic involutions giving rise to parallel orientations is investigated as a complementary language-theoretic concept.

## 2.3 Examples of Hairpin Languages

Consider the following examples of hairpin languages:

(a) Let $k \geq 1$ and

$$S_k = \left\{ \binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^* \right\} \binom{f}{f}^* \binom{b}{b}^{\geq k} \binom{f}{f}^* \left\{ \varepsilon, \binom{f}{\varepsilon} \right\}. \tag{2}$$

The general form of the DNA bonds specified by $S_k$ (when $\theta$ is an antimorphism) is represented by Figure 3. That is, when $\theta$ is an antimorphism, only one bonded region (the *stem*) is formed in this simple hairpin structure, and the length of this stem is at least $k$. The set $shpf_{\Sigma}(S_k,\theta)$ is the set of all $\theta$-$k$-hairpin-free words, studied by Kari *et al.* [14].

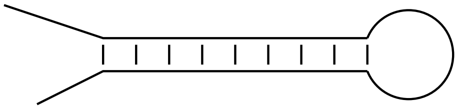

Figure 3: A simple hairpin structure.

(b) Let $k, m_1, m_2 \geq 1$. Jonoska *et al.* [10, 11] define $\theta(k,m_1,m_2)$-subword compliant languages, which are characterized by the following set of trajectories $S_{k,m_1,m_2}$:

$$S_{k,m_1,m_2} = \left( \binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^* \right) \binom{f}{f}^* \binom{b}{b}^{\geq k} \left( \bigcup_{m=m_1}^{m_2} \binom{f}{\varepsilon}^m \right).$$

In particular, a language $L \subseteq \Sigma^*$ is $\theta(k,m_1,m_2)$-subword compliant for a morphic or antimorphic involution $\theta$ if $L \subseteq shpf_{\Sigma}(S_{k,m_1,m_2},\theta)$.

(c) Let $k \geq 1$ and $S_k$ be defined by

$$S_k = \left(\left(\left(\binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^*\right)\binom{f}{f}^*\binom{b}{b}\right)^{\geq k}\binom{f}{f}^*\left\{\varepsilon, \binom{f}{\varepsilon}\right\}. \tag{3}$$

The shape described by this set of trajectories is called *scattered hairpins* by Kari *et al.* [12]. In particular, the condition is equivalent to the following: $x \leq_e u$ and $\theta(x) \leq_e v$ imply $|x| < k$. An example of the shape of scattered hairpins described by $S_k$ when $\theta$ is an antimorphism is given in Figure 4. The set $shpf_\Sigma(S_k, \theta)$ is denoted by $shpf(\theta, k)$ by Kari *et al.* [12].

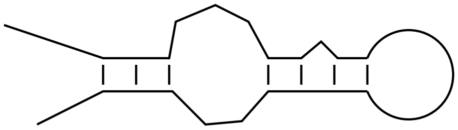

Figure 4: A scattered hairpin structure.

By adding DNA trajectories to scattered hairpins, we can also define familiar languages which have been studied by researchers in formal language theory. We begin by demonstrating that the classical languages of palindromes (modulo short palindromes) and squares are definable by a trajectory-based hairpin condition:

**Example 2.5.** Let $S_p = \binom{b}{b}^*\{\varepsilon, \binom{f}{\varepsilon}\}$. Then $shp_\Sigma(S_p, \mu) = \{x \in \Sigma^* \ : |x| \geq 2, x = x^R\}$.

To see this, note that the scattered hairpin conditions states that if $w \in shp_\Sigma(S, \mu)$, then there exists a factorization $w = uv$, $s \in \binom{b}{b}^*\{\varepsilon, \binom{f}{\varepsilon}\}$, and a word $x$, with $|x| \geq 1$ such that

$$x \, \omega_{\varphi_u(s)} \, u, \text{ and } x^R \, \omega_{\varphi_d(s)^R} \, v.$$

Note that $\varphi_u(s) \in 0^* \cup 0^*1$ and $\varphi_d(s)^R \in 0^*$. Thus, we have that $w \in x(\{\varepsilon\} \cup \Sigma)x^R$. Therefore, $w$ is a palindrome. The reverse inclusion is easily established.

The following example is established in the same way:

**Example 2.6.** Let $S_s = \binom{b}{b}^*$. Then $shp_\Sigma(S_s, \iota) = \{xx : x \in \Sigma^+\}$.

# 3 Preliminary Results

We first consider the implications of choosing alternate definitions for hairpin-freeness using DNA trajectories. In the first case, we show that, with DNA trajectories, there is no increase in power by adding a parameter $k \geq 1$ which enforces a minimum length of the (scattered) stem of the hairpin. In the second case, we show that if separate DNA trajectories are allowed to be chosen for the bonding on both sides of the stem, the result can destroy the structure described by the set of DNA trajectories.

In particular, consider the following definition:

**Definition 3.1.** Let $k \geq 1$ and $S \subseteq V_D^*$. Say a word $w$ is $\theta$-$k$-$S$-hairpin-free (or $shp_\Sigma(S,\theta,k)$-free) if the following condition holds:

$$\forall u,v,x, \in \Sigma^*, s \in S, (w = uv, x\, \omega_{\varphi_u(s)}\, u, \theta(x)\, \omega_{\varphi_d(s)^R}\, v) \Rightarrow (|x| < k).$$

This definition more closely mirrors the definitions provided by Kari *et al.* [12, 14]. Let $shpf_\Sigma(S,\theta,k)$ denote the set of $shp_\Sigma(S,\theta,k)$-free words. We now show that sets of DNA trajectories are sufficiently powerful to eliminate the need for considering $S$-$\theta$-$k$-hairpin-free words.

**Lemma 3.2.** *Let $k \geq 1$ and $S \subseteq V_D^*$ be a set of DNA trajectories. There exists a set of DNA trajectories $S' \subseteq V_D^*$ such that $shpf_\Sigma(S,\theta,k) = shpf_\Sigma(S',\theta)$.*

*Proof.* Let $S'$ be defined by $S' = S - \{s \in V_D^* \, : \, |s|_{\binom{b}{b}} < k\}$. Let $z \notin shpf_\Sigma(S,\theta,k)$. Then there exist $u,v,x \in \Sigma^*, s \in S$ with $z = uv$ and $|x| \geq k$ such that

$$x\, \omega_{\varphi_u(s)}\, u \text{ and } \theta(x)\, \omega_{\varphi_d(s)^R}\, v.$$

Note that $|x| = |\varphi_u(s)|_0 = |s|_{\binom{b}{b}}$. Thus, $|s|_{\binom{b}{b}} \geq k$ and $s \in S'$. Therefore, $z \notin shpf_\Sigma(S',\theta)$.

Similarly, if $z \notin shpf_\Sigma(S',\theta)$ then there exist $u,v,w \in \Sigma^*$, $s \in S'$ with $z = uv$ and $w \neq \varepsilon$ such that

$$w\, \omega_{\varphi_u(s)}\, u \text{ and } \theta(w)\, \omega_{\varphi_d(s)^R}\, v.$$

Note that $|s|_{\binom{b}{b}} = |\varphi_u(s)|_0 \geq k$ as $s \in S'$. Thus, $|w| \geq k$ and $z \notin shpf_\Sigma(S,\theta,k)$. □

For fixed $k$, the construction in Lemma 3.2 does not alter the complexity of $S$ if $S$ lies in a language class which is closed under finite modification[1].

We also consider the implications of choosing a single DNA trajectory in the definition of hairpin-freeness. In particular, note that a single DNA trajectory $s$ is used in both $\omega_{\varphi_u(s)}$ and $\omega_{\varphi_d(s)^R}$ in the definition (1). This reflects that a single DNA trajectory defines the bonding on both sides of the (perhaps scattered) stem of the hairpin. If separate $s_1, s_2 \in S$ are allowed to be chosen, i.e., using $\omega_{\varphi_u(s_1)}$ and $\omega_{\varphi_d(s_2)^R}$, then the structure of the set $S$ can be destroyed. For example, consider the set

$$S = \binom{f}{\varepsilon}^* \binom{b}{b}^+ \binom{f}{f}^* \left\{\varepsilon, \binom{f}{\varepsilon}\right\} \cup \binom{\varepsilon}{f}^* \binom{b}{b}^+ \binom{f}{f}^* \left\{\varepsilon, \binom{f}{\varepsilon}\right\}. \qquad (4)$$

In the case of an antimorphic involution, $S$ is represented graphically in Figure 5. Note that $\varphi_u(S) = 1^*0^+1^*$ while $\varphi_d(S) = 1^*0^+1^*$. Thus, if separate $s_1, s_2 \in S$ are chosen, the possibility of choosing $s_1, s_2$ with $\varphi_u(s_1) = 1^{\ell_1}0^i1^j$ and $\varphi_d(s_2) = 1^{\ell_2}0^i1^k$ destroys the bonding described in (4) and depicted in Figure 5, as these choices of $s_1, s_2$ also forbid hairpins of the form

$$\left\{\binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^*\right\} \binom{f}{f}^* \binom{b}{b}^+ \binom{f}{f}^* \left\{\varepsilon, \binom{f}{\varepsilon}\right\}, \qquad (5)$$

depicted (in the case of an antimorphic involution) in Figure 6. The analogous observation for bond-free properties—that a single DNA trajectory should be used to define both the upper and lower bonding—is examined by the author [4].

[1] A language class $\mathscr{C}$ is closed under finite modification if for all $L \in \mathscr{C}$ and all words $x$, $L \cup \{x\}, L - \{x\} \in \mathscr{C}$. Most common language classes are closed under finite modification; an example of a class that is not is the class of 0L languages.

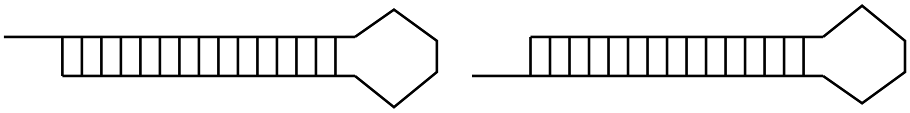

Figure 5: Graphical representation of (4).

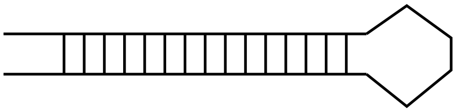

Figure 6: Graphical representation of (5).

# 4 Containment and Equivalence

We begin with some preliminary results on containment and equivalence between sets of DNA trajectories defining hairpin languages. These results are easily established, but are required in what follows.

**Proposition 4.1.** *Let $S_1, S_2 \subseteq V_D^*$ with $S_1 \subseteq S_2$, $\Sigma$ be an alphabet and $\theta : \Sigma^* \to \Sigma^*$ be a morphic or antimorphic involution. Then the following inclusion holds $shp_\Sigma(S_1, \theta) \subseteq shp_\Sigma(S_2, \theta)$.*

We also note that distinct trajectories may represent the same bonding pattern. For instance, note that an occurrence of $\binom{f}{f}$ is equivalent to an occurrence of $\binom{f}{\varepsilon}\binom{\varepsilon}{f}$. Due to this equivalence, we show the existence of a normal form for sets of DNA trajectories which is sometimes useful.

**Lemma 4.2.** *For all sets of DNA trajectories $S \subseteq V_D^*$ there exists a set of DNA trajectories $S' \subseteq \left(\left(\binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^*\right)\binom{f}{f}^*\binom{b}{b}^*\right)^*\binom{f}{f}^*\left\{\binom{f}{\varepsilon}, \varepsilon\right\}$ such that $shp_\Sigma(S, \theta) = shp_\Sigma(S', \theta)$.*

*Proof.* Consider the following rewriting rules:

$$\begin{array}{lcl@{\qquad\qquad}lcl}
\binom{f}{\varepsilon}\binom{f}{f} & \longleftrightarrow & \binom{f}{f}\binom{f}{\varepsilon} & \binom{\varepsilon}{f}\binom{f}{f} & \longleftrightarrow & \binom{f}{f}\binom{\varepsilon}{f} \\
\binom{f}{\varepsilon}\binom{\varepsilon}{f} & \longleftrightarrow & \binom{f}{f} & \binom{\varepsilon}{f}\binom{f}{\varepsilon} & \longleftrightarrow & \binom{f}{f}.
\end{array}$$

Clearly, none of the above rules alter the words which are $shp_\Sigma(S, \theta)$-free. Thus, to put $S$ in the required form, we simply migrate extra occurrences of $\binom{f}{\varepsilon}$ or $\binom{\varepsilon}{f}$ to the left-hand side of non-$\binom{b}{b}$ blocks. The loop section of the hairpin is the exception. We deal with this by observing that, for example, the block $\binom{f}{\varepsilon}^3$ interpreted as a loop is equivalent to $\binom{f}{f}\binom{f}{\varepsilon}$. □

If $S$ is in the form specified by Lemma 4.2, we say that $S$ is in *normal form*. Further, if $S \subseteq V_D^*$, then by $[S]$ we mean the set of all DNA trajectories which can be rewritten to a DNA trajectory $s \in S$ by using the above rules.

# 5 Closure Properties

In this section we examine the closure properties of hairpin languages based on the complexity of $S$. Examples 2.5 and 2.6 immediately yield the following lemma:

**Lemma 5.1.** *There exist a regular set of DNA trajectories $S$ and an antimorphic involution $\theta$ (resp., morphic involution $\sigma$) such that $shp_\Sigma(S,\theta)$ is not a regular language (resp., $shp_\Sigma(S,\sigma)$ is not a CFL).*

Note that Lemma 5.1 is in contrast to the case of hairpin languages and scattered hairpin languages, studied by Kari *et al.* [12], where the associated languages are regular. Despite the fact that regularity is not preserved when using a set of DNA trajectories to describe hairpin trajectories, we can show that for all regular sets of trajectories $S$ and all antimorphic involutions $\theta$, the language $shp_\Sigma(S,\theta)$ is always context-free:

**Theorem 5.2.** *If $\theta$ is an antimorphic involution and $S$ is a regular set of DNA trajectories, then $shp_\Sigma(S,\theta)$ is a linear context-free language.*

*Proof.* Let $S$ be a regular set of DNA trajectories and $G_S = (V_N, V_D, P_S, A_0)$ be a left-linear grammar for $S$, where all productions in $P_S$ are of the form $A \rightarrow aB$ or $A \rightarrow a$ for $a \in V_D$ and $A, B \in V_N$ (since we can assume without loss of generality that $\varepsilon \notin S$). Let $G = (V_N, \Sigma, P, A_0)$ be the CFG defined as follows: for all $A \rightarrow t\alpha$ in $P_S$, with $t \in V_D$ and $\alpha \in V_N \cup \{\varepsilon\}$, we add the following productions to $P$:

(a) if $t = \binom{f}{\varepsilon}$, the productions $A \rightarrow a\alpha$ are added for all $a \in \Sigma$.

(b) if $t = \binom{\varepsilon}{f}$, the productions $A \rightarrow \alpha a$ are added for all $a \in \Sigma$.

(c) if $t = \binom{f}{f}$, the productions $A \rightarrow a\alpha b$ are added for all $a, b \in \Sigma$.

(d) if $t = \binom{b}{b}$, the productions $A \rightarrow a\alpha\theta(a)$ are added for all $a \in \Sigma$.

To verify that this works, note that if $A_0 \Rightarrow^*_{G_S} s$, then in $G$ we can build any word $w$ which bonds according to $s$ from both the left and right ends of $w$. Since $\theta$ is an antimorphic involution, this process builds bonded regions which are oriented in the proper fashion. Thus, $L(G) = shp_\Sigma(S,\theta)$. □

We note that if we relax the condition that $S$ is regular, Theorem 5.2 does not hold.

**Lemma 5.3.** *Let $\Sigma$ be an alphabet with $|\Sigma| \geq 3$. There exists a (linear) context-free set of DNA trajectories $S \subseteq V_D^*$ such that $shp_\Sigma(S,\mu)$ is not a CFL.*

*Proof.* Let $S = \left\{ \binom{f}{\varepsilon}^n \binom{b}{b}^n \ : \ n \geq 0 \right\}$. Then we note that if $w \in shp_\Sigma(S,\mu)$, then there exist a factorization $w = uv$, $s \in S$ and a word $x$ such that

$$\begin{array}{ccc} x & \omega_{\varphi_u(s)} & u \\ x^R = \theta(x) & \omega_{\varphi_d(s)^R} & v. \end{array}$$

As $\varphi_d(s) \in 0^*$, the second relation implies $v = x^R$. Further, as $\varphi_u(s) \in \{1^n 0^n \ : \ n \geq 0\}$, the first relation implies $u = yx$ where $|y| = |x|$. Thus, we can verify the equality

$$shp_\Sigma(S,\mu) = \{yxx^R \ : \ x \in \Sigma^+, y \in \Sigma^*, |x| = |y|\}.$$

To see that $shp_\Sigma(S,\mu)$ is not a CFL, note that

$$shp_\Sigma(S,\mu) \cap c^+ba^+cca^+b = \{c^{n+2}ba^ncca^nb \ : \ n \geq 0\},$$

which is not a CFL by an application of the pumping lemma. □

Further, if we consider $shpf_\Sigma(S,\theta)$ for regular $S$ and antimorphic $\theta$ the result may not be context-free. Considering Lemma 5.1, we may expect this result, as the CFLs are not closed under complement.

**Theorem 5.4.** *Let $\Sigma$ be an alphabet with $|\Sigma| \geq 3$. There exist a regular set of DNA trajectories $S \subseteq V_D^*$ and an antimorphic involution $\theta$ such that $shpf_\Sigma(S,\theta)$ is not a CFL.*

*Proof.* Let $S_1 = \binom{\varepsilon}{f}^* \binom{b}{b}^+$, $S_2 = \binom{f}{\varepsilon}^* \binom{b}{b}^+$ and $S = S_1 \cup S_2$. Note that $S$ is regular. We claim that

$$shp_\Sigma(S,\mu) = \{ww^Rx \ : \ w \in \Sigma^+, x \in \Sigma^*\} \ \cup \ \{xww^R \ : \ w \in \Sigma^+, x \in \Sigma^*\}.$$

To see this equality, let $z \in shp_\Sigma(S,\mu)$. Then either $z \in shp_\Sigma(S_1,\mu)$ or $z \in shp_\Sigma(S_2,\mu)$. Consider the former case. For all $s \in S_1$, there exist $i, j \in \mathbb{N}$ with $j \geq 1$ such that $s = \binom{\varepsilon}{f}^i \binom{b}{b}^j$. Note that $\varphi_u(s) = 0^j$ and $\varphi_d(s)^R = 0^j 1^i$. Thus, as $z$ has an $S_1$-hairpin, there exist $u, v, x \in \Sigma^*$ such that $x \neq \varepsilon$, $z = uv$, $x \, \omega_{0^j} \, u$ and $x^R = \mu(x) \, \omega_{0^j 1^i} \, v$. This implies that $x = u$ and $v = x^R y$ for some $y \in \Sigma^*$. Thus, $z = uv = xx^R y$. If $z \in shp_\Sigma(S_2,\mu)$, we similarly get that $z = xww^R$ for some $x \in \Sigma^*, w \in \Sigma^+$. The reverse inclusion is proven similarly.

Thus, we have that

$$shpf_\Sigma(S,\mu) = \{x \ : \ \forall y \in \Sigma^+, z \in \Sigma^*, x \notin \{yy^Rz, zyy^R\}\}. \tag{6}$$

From (6), we can easily see that $shpf_\Sigma(S,\mu)$ is not context-free. In particular,

$$shpf_\Sigma(S,\mu) \cap ba^+cca^+bba^+c = \{ba^icca^jbba^kc \ : \ i,j,k \geq 1, i \neq j \text{ and } j \neq k\}.$$

By an application of the pumping lemma for CFLs, this language is not context-free, establishing the result. □

Thus, in general, $shpf_\Sigma(S,\theta)$ is not a CFL if $S$ is regular and $\theta$ is an antimorphism. However, we can find conditions on $S$ such that $shpf_\Sigma(S,\theta)$ is a CFL for all antimorphic involutions $\theta$. We require some additional notions.

For $s \in V_D^*$, we define the *yield length* of $s$, denoted $||s||$, by $||s|| = |\varphi_u(s)| + |\varphi_d(s)|$. Thus, if $x \in shp_\Sigma(\{s\},\theta)$ (for any choice of $\theta$) then $|x| = ||s||$. For $S \subseteq V_D^*$, we let $||S|| = \{||s|| \ : \ s \in S\} \subseteq \mathbb{N}$. The following technical lemma is easily established:

**Lemma 5.5.** *Let $S \subseteq V_D^*$ be a regular set of DNA trajectories. Then the language $\{x \in \Sigma^* \ : \ |x| \in ||S||\}$ is a regular language for all alphabets $\Sigma$.*

*Proof.* Consider the morphism $\rho : V_D^* \to \{b, f\}^*$ defined by $\rho(\binom{x}{y}) = xy$ for all $\binom{x}{y} \in V_D$. Then note that if $\tau : \{b, f\}^* \to 2^{\Sigma^*}$ is the finite substitution defined by $\tau(b) = \tau(f) = \Sigma$, then $\tau(\rho(S)) = \{x \in \Sigma^* \ : \ |x| \in ||S||\}$. The result follows by the closure of the regular languages under morphisms and finite substitutions. □

Recall the definition of the density function $p_L$ of a language $L$. Define $p_L : \mathbb{N} \to \mathbb{N}$ by $p_L(n) = |L \cap \Sigma^n|$ for all $n \geq 0$. That is, $p_L(n)$ gives the number of words of length $n$ in $L$. Call a language $L$ *slender* if $p_L(n) \in O(1)$ [19].

We can now demonstrate a nontrivial class of sets of DNA trajectories for which the set of hairpin-free words will be guaranteed to be a CFL:

**Theorem 5.6.** *Let $S \subseteq V_D^*$ be a slender regular set of DNA trajectories. Then for all antimorphic involutions $\theta$, $shpf_\Sigma(S, \theta)$ is a CFL.*

*Proof.* Since the CFLs are closed under union, it is enough to show that the result holds for regular sets of trajectories with density at most one.

Let $S$ be a regular set of DNA trajectories with density at most one and $G_S = (V_N, V_D, P_S, A_0)$ be a left-linear grammar for $S$, where all productions are of the form $A \to tB$ or $A \to t$ for $t \in V_D$ and $A, B \in V_N$ (again, we can assume $\varepsilon \notin S$). Let $\widehat{V_N}$ be a copy of $V_N$, and $G = (V_N \cup \widehat{V_N}, \Sigma, P, A_0)$ be the CFG defined as follows.

For all productions of the form $A \to tB$, where $A, B \in V_N$ and $t \in V_D$, we perform the following actions:

(a) If $t = \binom{f}{\varepsilon}$, then add to $P$ the productions $A \to aB$ and $\widehat{A} \to a\widehat{B}$ for all $a \in \Sigma$.

(b) If $t = \binom{\varepsilon}{f}$, then add to $P$ the productions $A \to Ba$ and $\widehat{A} \to \widehat{B}a$ for all $a \in \Sigma$.

(c) If $t = \binom{f}{f}$, then add to $P$ the productions $A \to aBb$ and $\widehat{A} \to a\widehat{B}b$ for all pairs $a, b \in \Sigma$.

(d) If $t = \binom{b}{b}$, then add to $P$ the productions

$$\begin{aligned} A &\to aB\theta(a) \quad \forall a \in \Sigma, && (7) \\ A &\to a\widehat{B}b \quad \forall a, b \in \Sigma, \theta(a) \neq b, && (8) \\ \widehat{A} &\to a\widehat{B}b \quad \forall a, b \in \Sigma. && (9) \end{aligned}$$

For all productions of the form $A \to t$, where $A \in V_N$ and $t \in V_D$, we perform the following actions:

(a) If $t = \binom{f}{\varepsilon}$ or $t = \binom{\varepsilon}{f}$, then add to $P$ the productions $\widehat{A} \to a$ for all $a \in \Sigma$.

(b) If $t = \binom{f}{f}$, then add to $P$ the productions $\widehat{A} \to ab$ for all pairs $a, b \in \Sigma$.

(c) If $t = \binom{b}{b}$, then add to $P$ the productions

$$\begin{aligned} A &\rightarrow ab \quad \forall a,b \in \Sigma, \theta(a) \neq b & (10)\\ \widehat{A} &\rightarrow ab \quad \forall a,b \in \Sigma. & (11) \end{aligned}$$

Note that the productions of $G$ are separated into two types: those whose left-hand side has a nonterminal from $V_N$, and those from $\widehat{V_N}$. Those from $V_N$ simulate $S$ much in the same way as the proof of Theorem 5.2, however, they are not allowed to be the final step of a derivation. To move to those which involve $\widehat{V_N}$, we must introduce a mismatch at some point where the trajectory sees $\binom{b}{b}$ (e.g., productions of the type (8) and (10)). Productions whose left-hand side is from $\widehat{V_N}$ are permitted to terminate a production. Further, as seen in (9) and (11), they are not constrained to match when encountering a $\binom{b}{b}$ in the trajectory—their only concern is to guarantee that the length of the derived word is equal to $||s||$ for some $s \in S$.

From this, we claim that $G$ generates the following language:

$$L(G) = shpf_{\Sigma}(S,\theta) \cap \{x \in \Sigma^* \ : \ |x| \in ||S||\}.$$

To see this, note that $L(G)$ is a subset of the left-hand side. For the reverse inclusion, if $x \in shpf_{\Sigma}(S,\theta) \cap \{x \in \Sigma^* \ : \ |x| \in ||S||\}$, then $|x| = ||s||$ where $s$ is the unique DNA trajectory in $S$ with length $|x|$. Note that our grammar will generate $x$ by ensuring that a mismatch is made in some position of $x$ where bonding is required to occur by $S$.

By Lemma 5.5, the language $\{x \in \Sigma^* \ : \ |x| \in ||S||\}$ is a regular language. Thus, its complement, $\{x \in \Sigma^* \ : \ |x| \notin ||S||\}$ is also a regular language, by the closure properties of the regular languages. We conclude that

$$shpf_{\Sigma}(S,\theta) = L(G) \cup \{x \in \Sigma^* \ : \ |x| \notin ||S||\}$$

is a CFL, by the closure properties of the context-free languages. □

Theorem 5.6 shows the power of using DNA trajectories for characterizing hairpins. By using a well-studied property of languages and applying it to the set of DNA trajectories, we can guarantee important properties of the associated hairpin language. However, in this case, we find that in addition to the complexity of the set of DNA trajectories, it is also another measure of the complexity—the density of the language—that yields the result.

Considering their role in Theorem 5.6, we can ask about the possible structure of slender regular sets of DNA trajectories. The following important result has been established independently by, e.g., Păun and Salomaa [19], Shallit [24] and, more generally, by Szilard *et al.* [25]:

**Theorem 5.7.** *A regular language $R$ over $\Sigma$ is slender if and only if there exist $k \geq 1$, and $x_i, y_i, z_i \in \Sigma^*$ for $1 \leq i \leq k$ such that $R = \bigcup_{i=1}^{k} x_i y_i^* z_i$.*

Thus, slender sets of DNA trajectories include the familiar case of palindromes from Example 2.5, but is not powerful enough, for example, to include the set of $\theta$-$k$-hairpin-free words studied by Kari *et al.* [14].

We now turn to the complexity of $shpf_{\Sigma}(S,\theta)$ for morphic involutions $\theta$. By Lemma 5.1, we know that $shp_{\Sigma}(S,\theta)$ can fail to be a CFL, even if $S$ is regular. However, the example given (Example 2.6) yields a language whose complement $shpf_{\Sigma}(S,\theta)$ *is* a CFL. However, we can find an example of a regular set $S$ such that $shpf_{\Sigma}(S,\theta)$ is not a CFL.

**Theorem 5.8.** *Let $\Sigma$ be an alphabet with $|\Sigma| \geq 3$. There exist a regular set of DNA trajectories $S \subseteq V_D^*$ and an morphic involution $\theta$ such that $shpf_\Sigma(S, \theta)$ is not a CFL.*

*Proof.* Let $S = \binom{\varepsilon}{f}^* \binom{b}{b}^* \cup \binom{f}{\varepsilon}^* \binom{b}{b}^*$. Then note that $shp_\Sigma(S, \iota) = \{xxw \ : \ x \in \Sigma^+, w \in \Sigma^*\} \cup \{wxx \ : \ x \in \Sigma^+, w \in \Sigma^*\}$. Thus, we get that $shpf_\Sigma(S, \iota) = \{x \in \Sigma^* \ : \ \forall y \in \Sigma^+, z \in \Sigma^*, x \notin \{yyz, zyy\}\}$. By intersecting with the regular language $(ba^+c)^3$, we note

$$shpf_\Sigma(S, \iota) \cap (ba^+c)^3 = \{ba^icba^jcba^kc \ : \ i \neq j \text{ and } j \neq k\},$$

which is not a CFL by an application of the pumping lemma. □

We now consider the complexity of $shp_\Sigma(S, \theta)$ for unary alphabets (i.e., $\Sigma$ with $|\Sigma| = 1$).

**Lemma 5.9.** *Let $|\Sigma| = 1$ and $\mathscr{L}$ be any class of languages closed under morphisms. If $S \subseteq V_D^*$ with $S \in \mathscr{L}$, then $shp_\Sigma(S, \theta) \in \mathscr{L}$ for all morphic and antimorphic involutions $\theta$.*

*Proof.* Note that if $|\Sigma| = 1$ and $\theta : \Sigma \to \Sigma$ is an involution, then when $\theta$ is extended to a morphism or antimorphism, the result is equivalent to applying the identity morphism. Thus, we may assume throughout that $\theta$ is the identity morphism.

Recall $\rho$ and $\tau$ from the proof of Lemma 5.5. First, in the case where $|\Sigma| = 1$, $\tau$ is a morphism. Further, note that $\{x \in \Sigma^* \ : \ |x| \in ||S||\} = shp_\Sigma(S, \theta) = \tau(\rho(S))$. Therefore, if $S \in \mathscr{L}$, then so is $\tau(\rho(S))$, by the assumed closure properties of $\mathscr{L}$ and the fact that $\rho$ and (in this case only) $\tau$ is a morphism. We conclude that $shp_\Sigma(S, \theta) \in \mathscr{L}$ and the result holds. □

As a corollary, we note that for unary alphabets, if $S$ is regular (resp., context-free) then $shp_\Sigma(S, \theta)$ is regular (resp., context-free). For context-free languages, this contrasts Lemma 5.3.

## 5.1 Regularity of Hairpin Languages

In the previous section, we have seen that for some regular set of DNA trajectories $S$ and antimorphic involution $\theta$, the associated hairpin language $shpf_\Sigma(S, \theta)$ is not context-free. If we restrict $S$ to be slender, then we can guarantee that $shpf_\Sigma(S, \theta)$ is context-free for all antimorphic involutions $\theta$. In this section, we consider tools which will allow us to establish that $shpf_\Sigma(S, \theta)$ (and $shp_\Sigma(S, \theta)$) is regular. Instead of further constraining $S$ by beginning with slender sets of DNA trajectories, we look at relations on $S$ that ensure regularity of $shpf_\Sigma(S, \theta)$.

We define a partial order $\prec$ on words over $V_D^*$. Let $s_1, s_2 \in V_D^*$ with

$$\varphi_u(s_1) = \prod_{i=1}^{n} 1^{j_i} 0^{k_i}, \text{ and } \varphi_d(s_1) = \prod_{i=1}^{n} 1^{\ell_i} 0^{k_i},$$

for $n \geq 0$ and $j_i, k_i, \ell_i \geq 0$ for all $1 \leq i \leq n$. Then $s_2 \prec s_1$ if there exist $\alpha_1, \ldots, \alpha_n \in \{0,1\}^*$ such that the following three conditions hold:

(i) $\varphi_u(s_2) = \prod_{i=1}^{n} 1^{j_i} \alpha_i$ and $\varphi_d(s_2) = \prod_{i=1}^{n} 1^{\ell_i} \alpha_i$;

(ii) $|\alpha_i| = k_i$ for all $1 \leq i \leq n$; and

(iii) $\prod_{i=1}^{n} \alpha_i \notin 1^*$.

We note that $\prec$ is also used to investigate bond-free properties between separate single strands of DNA [4].

The situation is illustrated (in the case of an antimorphic involution) in Figure 7. The figure illustrates that if $s_2 \prec s_1$, then we can get from $s_1$ to $s_2$ by replacing a bonding region of length $k_i$ in $s_1$ with a region which is not completely bonded, but still has length $k_i$, in $s_2$.

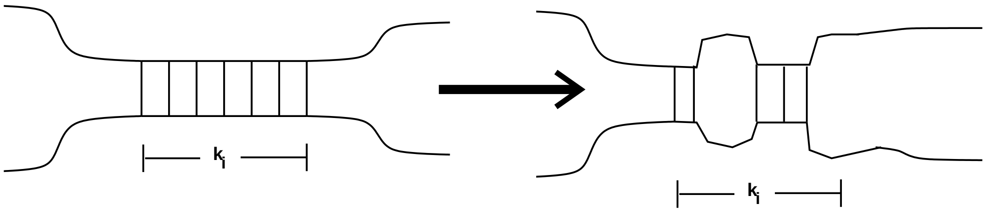

Figure 7: A portion of $s_1$ is shown on the left, and a portion of $s_2$ is shown on the right.

**Example 5.10.** Consider $s_1, s_2 \in V_D^*$ given by

$$s_1 = \begin{pmatrix} b \\ b \end{pmatrix}\begin{pmatrix} b \\ b \end{pmatrix}\begin{pmatrix} f \\ \varepsilon \end{pmatrix}\begin{pmatrix} b \\ b \end{pmatrix}\begin{pmatrix} f \\ f \end{pmatrix}, \; s_2 = \begin{pmatrix} b \\ b \end{pmatrix}\begin{pmatrix} f \\ f \end{pmatrix}\begin{pmatrix} f \\ \varepsilon \end{pmatrix}\begin{pmatrix} f \\ \varepsilon \end{pmatrix}\begin{pmatrix} f \\ f \end{pmatrix}\begin{pmatrix} \varepsilon \\ f \end{pmatrix}.$$

Note that $\varphi_u(s_1) = 00101, \varphi_u(s_2) = 01111, \varphi_d(s_1) = 0001,$ and $\varphi_d(s_2) = 0111$. Thus, $s_2 \prec s_1$ holds with $\alpha_1 = 01$ and $\alpha_2 = 1$.

Note that Example 5.10 demonstrates that the relation $\prec$ is not simply defined by the idea “possibly replace $\binom{b}{b}$ with $\binom{f}{f}$”. This is due to the equivalence of trajectories seen in the normal form of Lemma 4.2. However, the replacement intuition is formalized in the following result.

**Proposition 5.11.** *Let $\pi : V_D^* \to 2^{V_D^*}$ be the substitution defined by $\pi(x) = x$ if $x \neq \binom{b}{b}$ and $\pi(\binom{b}{b}) = \{\binom{b}{b}, \binom{f}{f}\}$. Then for all $S \subseteq V_D^*$,*

$$\left[\pi(S) \cap (V_D^* \begin{pmatrix} b \\ b \end{pmatrix} V_D^*)\right] = \{x \in V_D^* \;:\; \exists s \in S, x \prec s\}.$$

We now define the minimal set of DNA trajectories with respect to $\prec$. For all $S \subseteq V_D^*$, let $\min(S) = \{s \in S \;:\; \forall t (\neq s) \in S, t \not\prec s\}$.

**Example 5.12.** Consider the $k$-hairpin languages (2):

$$S_k = \left\{ \begin{pmatrix} f \\ \varepsilon \end{pmatrix}^* \cup \begin{pmatrix} \varepsilon \\ f \end{pmatrix}^* \right\} \begin{pmatrix} f \\ f \end{pmatrix}^* \begin{pmatrix} b \\ b \end{pmatrix}^{\geq k} \begin{pmatrix} f \\ f \end{pmatrix}^* \left\{ \varepsilon, \begin{pmatrix} f \\ \varepsilon \end{pmatrix} \right\}.$$

Note that if we put $\min(S_k)$ in normal form, we get

$$\min(S_k) = \left\{ \begin{pmatrix} f \\ \varepsilon \end{pmatrix}^* \cup \begin{pmatrix} \varepsilon \\ f \end{pmatrix}^* \right\} \begin{pmatrix} f \\ f \end{pmatrix}^* \begin{pmatrix} b \\ b \end{pmatrix}^{k} \begin{pmatrix} f \\ f \end{pmatrix}^* \left\{ \varepsilon, \begin{pmatrix} f \\ \varepsilon \end{pmatrix} \right\}.$$

We now show that $S$ and $\min(S)$ describe the same hairpin languages:

**Theorem 5.13.** *Let $S \subseteq V_D^*$. For all $\Sigma$ and all morphic or antimorphic involutions $\theta$, we have $shp_\Sigma(S,\theta) = shp_\Sigma(min(S),\theta)$.*

*Proof.* The inclusion $shp_\Sigma(S,\theta) \supseteq shp_\Sigma(\min(S),\theta)$ is immediate, by Lemma 4.1. Thus, let $x \in shp_\Sigma(S,\theta)$. Thus, there exist $u,v,w \in \Sigma^*$ with $w \neq \varepsilon$ and $s \in S$ such that $x = uv$, $w\, \omega_{\varphi_u(s)}\, u$ and $\theta(w)\, \omega_{\varphi_d(s)^R}\, v$. Assume $s \notin \min(S)$. Then there must exist $s' \in \min(S)$ such that $s' \prec s$. Let $n, \alpha_i, j_i, k_i, \ell_i$ (with $1 \leq i \leq n$) be defined so that $\varphi_u(s) = \prod_{i=1}^n 1^{j_1} 0^{k_i}, \varphi_d(s) = \prod_{i=1}^n 1^{\ell_i} 0^{k_i}, \varphi_u(s') = \prod_{i=1}^n 1^{j_1} \alpha_i$ and $\varphi_d(s') = \prod_{i=1}^n 1^{\ell_i} \alpha_i$. Let $\alpha = \prod_{i=1}^n \alpha_i$.

Consider that $|w| = \sum_{i=1}^n k_i$. Let $m = \sum_{i=1}^n k_i$, $w = \prod_{i=1}^m w_i$ where $w_i \in \Sigma$.

Further, let $\alpha = \prod_{i=1}^m \beta_i$ where $\beta_i \in \{0,1\}$. Define $I \subseteq \mathbb{N}$ by $I = \{i \ :\ \beta_i = 0\}$, i.e., exactly those positions of $\alpha$ which are zero. We want to consider those positions of $w$ which correspond to indices in $I$, since these correspond to the portion of $x$ which will remain bonded when we pass from $s$ to $s'$. Thus, let $w' = \prod_{i \in I} w_i$. From this definition, it is not hard to see that

$$w'\, \omega_{\varphi_u(s')}\, u \text{ and } \theta(w')\, \omega_{\varphi_d(s')^R}\, v.$$

As $\alpha \notin 1^*$, note that $I \neq \emptyset$ and so $w' \neq \varepsilon$. As $x = uv$, we have $x \in shp_\Sigma(\min(S),\theta)$. □

**Example 5.14.** Continuing with Example 5.12, we see that $shp_\Sigma(\min(S_k),\theta)$ (and thus $shp_\Sigma(S_k,\theta)$) is regular for all morphic or antimorphic involutions $\theta$. This was first established by Kari *et al.* [14, Prop. 3].

To see that $shp_\Sigma(\min(S_k),\theta)$ is regular, note that there are only finitely many occurrences of $\binom{b}{b}$ in each trajectory in $\min(S_k)$, and further that the two blocks of non-$\binom{b}{b}$s in each trajectory does not define a length restriction between distinct portions of the word $w \in shp_\Sigma(\min(S_k),\theta)$ – in particular, if $w = uxv\theta(x)y \in shp_\Sigma(\min(S_k),\theta)$, then there is no relationship between $u$ and $y$ or between $|u|$ and $|y|$. Thus, a finite automaton can verify if $w \in shp_\Sigma(\min(S_k),\theta)$ by storing finitely many symbols of $w$ in its finite control (those that correspond to the bonded portion $x$).

**Example 5.15.** The $k$-scattered hairpin languages (see (3)) are given by

$$S_k = \left( \left\{ \binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^* \right\} \binom{f}{f}^* \binom{b}{b} \right)^{\geq k} \binom{f}{f}^* \left\{ \varepsilon, \binom{f}{\varepsilon} \right\}.$$

Note that if $\min(S_k) \subseteq S_k$ is put in normal form, we get

$$\min(S_k) = \left( \left\{ \binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^* \right\} \binom{f}{f}^* \binom{b}{b} \right)^{k} \binom{f}{f}^* \left\{ \varepsilon, \binom{f}{\varepsilon} \right\}.$$

We can again establish that $shp_\Sigma(\min(S_k),\theta)$ is regular for all morphic or antimorphic involutions and all $k \geq 1$, with an argument similar to that in Example 5.14. Thus, as $shp_\Sigma(\min(S_k),\theta)$ is regular, so is $shp_\Sigma(S_k,\theta)$ (this was established by Kari *et al.* [12, Prop. 13(ii)]).

## 5.2 Finiteness of Hairpin Classes

We continue our investigation of conditions on $S$ and $\theta$ that ensure that $shp_{\Sigma}(S,\theta)$ and $shpf_{\Sigma}(S,\theta)$ lie within a certain class of languages by considering conditions on $S$ to ensure that $shpf_{\Sigma}(S,\theta)$ is finite. Kari *et al.* [14] have studied conditions which ensure finiteness of hairpin-free languages. We summarize their results here:

**Theorem 5.16.** *Let $k \geq 1$ and $S_k \subseteq V_D^*$ be defined by*

$$S_k = \left\{ \binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^* \right\} \binom{f}{f}^* \binom{b}{b}^{\geq k} \binom{f}{f}^* \left\{ \binom{f}{\varepsilon}, \varepsilon \right\}.$$

*Then $shpf_{\Sigma}(S,\theta)$ is finite if and only if*

*(i)* $\theta = \iota$,

*(ii)* $\theta = \mu$ *and* $k = 1$, *or*

*(iii)* $\theta = \mu$, $|\Sigma| = 2$ *and* $k \leq 4$.

We note that Rampersad and Shallit [20] have independently established stronger results than Theorem 5.16 (ii) and (iii): they show that the same languages are finite even if we allow the occurrence of a word and the occurrence of its reversal (i.e., both single-stranded portions of the stem) to overlap. However, the constructions of Kari *et al.* and Rampersad and Shallit are essentially the same.

As noted by Kari *et al.* [14], problems concerning finiteness of hairpin languages are occasionally related to problems in combinatorics on words (see Choffrut and Karhumäki [2] or Lothaire [16] for an introduction to combinatorics on words). We recall some notions of avoidability. Let $V$ be a set of indeterminates. Then we say that a word $w \in \Sigma^*$ *encounters* the pattern $\alpha \in V^*$ if there exists a morphism $h : V^* \to \Sigma^*$ with $h(\beta) \neq \varepsilon$ for all $\beta \in V$ (i.e., $h$ is *non-erasing*) such that $h(\alpha)$ is a subword of $w$. We say that $w$ *avoids* $\alpha$ if $w$ does not encounter $\alpha$.

We say that a pattern $\alpha$ is *avoidable* if there exists arbitrarily long words which avoid $\alpha$. We say that $\alpha$ is $k$-avoidable if there exists arbitrarily long words $w \in \Sigma^*$, where $|\Sigma| = k$, such that $w$ avoids $\alpha$. The terms unavoidable and $k$-unavoidable are defined in the natural way.

Using results from the study of combinatorics on words, we can instantly conclude the finiteness of some scattered-hairpin languages by virtue of their coinciding with known unavoidable patterns. In particular, we use the results of Cassaigne [1], who gives a list of avoidability of patterns over 2- and 3-letter pattern alphabets, to derive finiteness results. These results are limited to the case where $\theta = \iota$, due to the emphasis on repetition of subwords in the study of combinatorics on words.

As an example, every sufficiently long word over any alphabet contains two occurrences of some letter. In terms of hairpins, we can phrase this equivalently as follows: the language $shpf_{\Sigma}(S,\iota)$ is finite for all $\Sigma$, where

$$S = \left\{ \binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^* \right\} \binom{f}{f}^* \binom{b}{b} \binom{f}{f}^* \left\{ \binom{f}{\varepsilon}, \varepsilon \right\}.$$

Using the tools of avoidability, we can also conclude the following:

**Lemma 5.17.** *Let* $S_1 = (\binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^*)\binom{f}{f}^*\binom{b}{b}^+\binom{f}{\varepsilon}^+\binom{b}{b}^+$. *The languages* $shpf_\Sigma(S_1, \iota)$ *are finite for all* $\Sigma$ *with* $|\Sigma| \leq 2$.

*Similarly, if* $S_2 = (\binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^*)\binom{f}{f}^*\binom{b}{b}^+\binom{\varepsilon}{f}^+\binom{b}{b}^+$, *the languages* $shpf_\Sigma(S_2, \iota)$ *are finite for all* $\Sigma$ *with* $|\Sigma| \leq 2$.

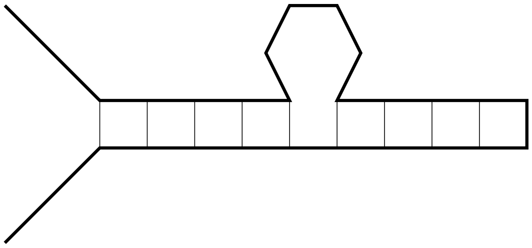

Figure 8: Hairpins described by $S_1$ in Lemma 5.17.

*Proof.* The set of trajectories $S_1$ describes the hairpin structure given in Figure 8. In particular, if $z$ is $shp_\Sigma(S_1, \iota)$-free, then $z$ is not of the form $uvxyyvw$ for some $v, x, y \in \Sigma^+$ and $u, w \in \Sigma^*$. Thus, $z$ avoids the pattern $ABCCA$. By Cassaigne [1, p. 157], this pattern is 2-avoidable. □

On the other hand, we can interpret the classic result of Entringer *et al.* [6] on avoidability of long squares in terms of hairpins:

**Theorem 5.18.** *Let* $S \subseteq V_D^*$ *be defined by* $S = \left\{ \binom{f}{\varepsilon}^* \cup \binom{\varepsilon}{f}^* \right\} \binom{f}{f}^* \binom{b}{b}^{\geq 3}$. *Then the language* $shpf_\Sigma(S, \iota)$ *is infinite if* $|\Sigma| \geq 2$.

For related results, see Fraenkel and Simpson [7] or Rampersad *et al.* [21]. Of course, there are both well-studied and novel problems in combinatorics on words and avoidability which cannot be expressed in terms of hairpins. For example, the avoidability of the pattern $XXX$, which is well-studied, cannot be expressed in terms of hairpins. However, the interaction between classical avoidability problems and hairpins is compelling, and the expressive power of hairpins suggests many problems, likely difficult, involving avoidability of patterns.

# 6 Decidability

We can now investigate the decidability of hairpin properties.

**Theorem 6.1.** *Given an antimorphism* $\theta$, *a regular set of DNA trajectories* $S$ *and a regular language* $L$, *it is decidable whether* $L$ *is* $shp_\Sigma(S, \theta)$*-free.*

*Proof.* Note that $L$ is $shp_\Sigma(S, \theta)$-free if and only if $L \cap shp_\Sigma(S, \theta) = \emptyset$. As $L$ is regular and $shp_\Sigma(S, \theta)$ is a CFL (by Theorem 5.2), it is decidable whether $L \cap shp_\Sigma(S, \theta) = \emptyset$. □

For undecidability, we show that there exists a regular set of trajectories $S$ such that determining whether context-free languages are $shp_\Sigma(S,\theta)$-free for morphic or antimorphic involutions is impossible. The following results employ the undecidability of Post's Correspondence Problem (PCP); we refer the reader to Harju and Karhumäki [8] for an introduction.

**Theorem 6.2.** *There exists a fixed regular set of DNA trajectories S such that the following problem is undecidable: "Given an alphabet $\Sigma$, an antimorphic involution $\theta : \Sigma^* \to \Sigma^*$ (resp., a morphic involution $\theta : \Sigma^* \to \Sigma^*$) and a CFL $L \subseteq \Sigma^*$, is $L \subseteq shpf_\Sigma(S,\theta)$?"*

*Proof.* We first consider the result for antimorphic involutions $\theta$. Let $S = \binom{b}{b}^* \binom{f}{\varepsilon}$ and let $I = (u_1, u_2, \ldots, u_n; v_1, v_2, \ldots, v_n)$ be a PCP instance over an alphabet $\Delta$. Let $\Sigma$ be the alphabet $\Sigma = \Delta \cup \{0, 1, \#, \overline{\#}\}$. Let $\theta : \Sigma^* \to \Sigma$ be the antimorphic involution defined by $\theta(a) = a$ for all $a \in \Sigma - \{\#, \overline{\#}\}$ and $\theta(\#) = \overline{\#}$. From the PCP instance $I$, we construct a context-free language $L$ via the grammar $G = (\{A_0, A_1, A_2\}, \Sigma, P, A_0)$, given by the following set of productions $P$:

$$\begin{array}{lcl} A_0 & \to & A_1\#A_2 \\ A_1 & \to & u_iA_10^i1 \quad \forall 1 \leq i \leq n \\ A_1 & \to & \varepsilon \\ A_2 & \to & 10^iA_2v_i^R \quad \forall 1 \leq i \leq n \\ A_2 & \to & \varepsilon \end{array}$$

The result will follow by the claim below:

**Claim 6.3.** *$L \cap shp_\Sigma(S,\theta) \neq \emptyset$ if and only if $I$ has a solution.*

($\Rightarrow$): Suppose a solution to $I$ is given by $x = u_{i_1}u_{i_2}\cdots u_{i_m} = v_{i_1}v_{i_2}\cdots v_{i_m}$ where $1 \leq i_j \leq n$ for $1 \leq j \leq m$. Consider the word

$$y = u_{i_1}u_{i_2}\cdots u_{i_m}0^{i_m}10^{i_{m-1}}1\cdots 0^{i_2}10^{i_1}1\#10^{i_1}10^{i_2}1\cdots 0^{i_{m-1}}10^{i_m}v_{i_m}^Rv_{i_{m-1}}^R\cdots v_{i_1}^R$$

Clearly, $y \in L$. Further, $y \in shp_\Sigma(S,\theta)$ via the DNA trajectory

$$s = \begin{pmatrix} b \\ b \end{pmatrix}^{|x|+m+\sum_{j=1}^m i_j} \begin{pmatrix} f \\ \varepsilon \end{pmatrix}.$$

($\Leftarrow$): Let $x \in L \cap shp_\Sigma(S,\theta)$. Thus, by the structure of $L$, we have

$$x = u_{i_1}u_{i_2}\cdots u_{i_m}0^{i_m}10^{i_{m-1}}1\cdots 0^{i_2}10^{i_1}1\#10^{\ell_1}10^{\ell_2}1\cdots 0^{\ell_{r-1}}10^{\ell_r}v_{\ell_r}^Rv_{\ell_{r-1}}^R\cdots v_{\ell_1}$$

with $1 \leq i_j, \ell_k \leq n$ for all $1 \leq j \leq m$ and $1 \leq k \leq r$. Consider that # appears in $x$, but $\overline{\#}$ does not. Thus, the occurrence of # must be unbonded, as $x \in shp_\Sigma(S,\theta)$, and each $s \in S$ only has one occurrence of $f$. Thus, we must have that if $x \in shp_\Sigma(S,\theta)$ via the DNA trajectory $s$, then $s = \binom{b}{b}^\alpha \binom{f}{\varepsilon}$, where $\alpha = |u_{i_1}u_{i_2}\cdots u_{i_m}0^{i_m}10^{i_{m-1}}1\cdots 0^{i_2}10^{i_1}1| = |10^{\ell_1}10^{\ell_2}1\cdots 0^{\ell_{r-1}}10^{\ell_r}v_{\ell_r}^Rv_{\ell_{r-1}}^R\cdots v_{\ell_1}|$

Note that $\theta(u_{i_1}\cdots u_{i_m}0^{i_m}1\cdots 0^{i_1}1) = 10^{i_1}\cdots 10^{i_m}u_{i_m}^Ru_{i_{m-1}}^R\cdots u_{i_1}^R$. As the alphabets $\Sigma$ and $\{0,1\}$ are disjoint, we must have

$$u_{i_1}u_{i_2}\cdots u_{i_m} = v_{\ell_1}v_{\ell_2}\cdots v_{\ell_r}$$

and $0^{i_1}10^{i_2}1\cdots0^{i_m}1 = 0^{\ell_1}10^{\ell_2}1\cdots0^{\ell_{r-1}}10^{\ell_r}1$. Therefore, $r = m$ and $i_j = \ell_j$ for all $1 \le j \le m$. Thus, $u_{i_1}\cdots u_{i_m} = v_{i_1}\cdots v_{i_m}$ represents a solution to the PCP instance.

The proof for the case morphic involutions is essentially the same as the proof above. In particular, given a PCP instance $I$ over an alphabet $\Delta$, the alphabet $\Sigma$ remains the same, the involution $\theta$ remains the same, but is extended to an morphism, and the set $S$ is also the same.

The change is that we define context-free language $L$ by the grammar $G = (\{A_0, A_1, A_2\}, \Sigma, P, A_0)$, given by the following set of productions $P$:

$$\begin{array}{lcll} A_0 & \to & A_1\#A_2 & \\ A_1 & \to & u_iA_10^i1 & \forall 1 \le i \le n \\ A_1 & \to & \varepsilon & \\ A_2 & \to & v_iA_20^i1 & \forall 1 \le i \le n \\ A_2 & \to & \varepsilon & \end{array}$$

In this case, we leave it to the reader to establish the claim that $L \cap shp_\Sigma(S,\theta) \neq \emptyset$ if and only if $I$ has a solution. □

Note that the key concept in Theorem 6.2 is that we inserted the symbol #, whose image $\overline{\#}$ (under $\theta$) did not appear in the language. In this way, we ensured that bonding did not occur at a specified position in our words.

# 7 Conclusions

In this paper, we have given a technique for modelling hairpin conditions on DNA words by using DNA trajectories. We have investigated closure properties and decidability questions relating to these hairpin sets. In order to ensure positive closure properties, restrictions must be placed on the sets of DNA trajectories. In particular, if $S$ is a slender regular set of DNA trajectories, then $shpf_\Sigma(S,\theta)$ is a context-free language for antimorphic involutions $\theta$. On the other hand, for all regular sets of DNA trajectories $S$ and all antimorphic involutions $\theta$, the set $shp_\Sigma(S,\theta)$ is a context-free language. In proving regularity of scattered hairpin sets, we have considered a partial order $\prec$ and the minimal set of DNA trajectories with respect to $\prec$.

With respect to decidability, we have shown that hairpin-freeness of a regular language is decidable for regular set of trajectories and antimorphic involutions. However, there exists a fixed regular set of trajectories $S$ such that it is undecidable, given an antimorphic involution and a context-free language $L$, whether or not $L$ is $shp_\Sigma(S,\theta)$-free.

One restriction on using DNA trajectories is that the model has the potential to be too precise: sets of DNA trajectories where bonds are enforced at particular positions are not realistic biological model, and would likely not be useful in DNA computing situations. Thus, care has to be taken in the choice of the set of DNA trajectories. The common choices for describing hairpin shapes in previous research all do not enforce strong conditions which are unrealistic; viewed as DNA trajectories, we note that the specifications are all infinite (regular) languages whose broad structure does not impose impossible conditions. A topic for future work is an investigation of the limitations of the use of DNA trajectories in modelling hairpins. In particular, we can impose a change to the definition (for instance, a probabilistic or similar model) which essentially ignores unrealistic conditions specified by a set of DNA trajectories, or use language theory to consider classes of sets of DNA trajectories which model realistic conditions and investigate their language theoretic properties.

Another topic for future research is the interplay between required formations and forbidden formations. Currently, given two sets of DNA trajectories $S_1$ and $S_2$, the sets $shp_\Sigma(S_1,\theta)$ and $shpf_\Sigma(S_2\theta)$ are independent entities. It might be a worthwhile extension to consider conditions which model statements such as "*hairpins from $S_1$ if necessary, but never from $S_2$*".

We feel that DNA trajectories are an appropriate and convenient tool for modelling hairpin conditions on words. The use of DNA trajectories also suggests interesting problems for further study, including further research on avoidability of patterns defined by scattered hairpin conditions.

# Acknowledgments

We thank the referees for their valuable comments which improved the presentation of this paper.

# References

[1] CASSAIGNE, J. *Motifs évitables et régularités dans les mots*. PhD thesis, Université Paris 6, 1994.

[2] CHOFFRUT, C., AND KARHUMÄKI, J. Combinatorics on words. pp. 329–438. In [23].

[3] DOMARATZKI, M. Trajectory-based embedding relations. *Fund. Inf. 59*, 4 (2004), 349–363.

[4] DOMARATZKI, M. Characterizing DNA bond shapes using trajectories. In *Developments in Language Theory* (2006), O. Ibarra and Z. Dang, Eds., vol. 4036 of *LNCS*, Springer, pp. 180–191.

[5] DOMARATZKI, M. Hairpin Structures Defined by DNA trajectories In *DNA 12* (2006), C. Mao and T. Yokomori, Eds., vol. 4287 of *LNCS*, Springer, pp. 182–194.

[6] ENTRINGER, R., JACKSON, D., AND SCHATZ, J. On nonrepetitive sequences. *J. Combin. Theory. Ser. A 16* (1974), 159–164.

[7] FRAENKEL, A., AND SIMPSON, J. How many squares must a binary sequence contain? *Electon. J. Combin. 2* (1995), #R2.

[8] HARJU, T., AND KARHUMÄKI, J. Morphisms. pp. 439–510. In [23].

[9] HOPCROFT, J. E., AND ULLMAN, J. D. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 1979.

[10] JONOSKA, N., KEPHART, D., AND MAHALINGAM, K. Generating DNA code words. *Congressus Numerantium 156* (2002), 99–110.

[11] JONOSKA, N., AND MAHALINGAM, K. Languages of DNA based code words. In *DNA Computing, 9th International Workshop on DNA Based Computers* (2004), J. Chen and J. Reif, Eds., vol. 2943 of *LNCS*, Springer, pp. 61–73.

[12] KARI, L., KONSTANTINIDIS, S., LOSSEVA, E., SOSÍK, P., AND THIERRIN, G. Hairpin structures in DNA words. In *DNA Computing* (2006), A. Carbone and N. Pierce, Eds., vol. 3892 of *Lecture Notes in Computer Science*, Springer, pp. 158–170.

[13] KARI, L., KONSTANTINIDIS, S., AND SOSÍK, P. On properties of bond-free DNA languages. *Theor. Comp. Sci. 334* (2005), 131–159.

[14] KARI, L., KONSTANTINIDIS, S., SOSÍK, P., AND THIERRIN, G. On hairpin-free words and languages. In *Developments in Language Theory: 9th International Conference* (2005), C. D. Felice and A. Restivo, Eds., vol. 3572 of *Lecture Notes in Computer Science*, Springer, pp. 296–307.

[15] KARI, L., LOSSEVA, E. KONSTANTINIDIS, S., SOSÍK, P., AND THIERRIN, G. A formal language analysis of DNA Hairpin Structures. *Fund. Inf. 71* (2006) 453–475.

[16] LOTHAIRE, M. *Combinatorics on Words*. Addison-Wesley, 1983.

[17] MATEESCU, A., ROZENBERG, G., AND SALOMAA, A. Shuffle on trajectories: Syntactic constraints. *Theor. Comp. Sci. 197* (1998), 1–56.

[18] PĂUN, G., ROZENBERG, G., AND SALOMAA, A. *DNA Computing: New Computing Paradigms*. Springer, 1998.

[19] PĂUN, G., AND SALOMAA, A. Thin and slender languages. *Disc. Appl. Math. 61* (1995), 257–270.

[20] RAMPERSAD, N., AND SHALLIT, J. Words avoiding reversed subwords. *J. Combin. Math. and Combin. Comput. 54* (2005), 157–164.

[21] RAMPERSAD, N., SHALLIT, J., AND WANG, M.-W. Avoiding large squares in infinite binary words. *Theor. Comp. Sci. 339* (2005), 19–34.

[22] ROTHEMUND, P., PAPADAKIS, N., AND WINFREE, E. Algorithmic self-assembly of DNA Sierpinski triangles. *PLoS Biol. 2*, 12 (2004), e424.

[23] ROZENBERG, G., AND SALOMAA, A., Eds. *Handbook of Formal Languages*. Springer, 1997.

[24] SHALLIT, J. Numeration systems, linear recurrences, and regular sets. *Inf. and Comp. 113*, 2 (1994), 331–347.

[25] Szilard, A., Yu, S., Zhang, K., and Shallit, J. Characterizing regular languages with polynomial densities. In *Mathematical Foundations of Computer Science 1992* (1992), I. Havel and V. Koubek, Eds., vol. 629 of *Lecture Notes in Computer Science*, Springer, pp. 494–503.