Pattern Avoidability with Involution

Bastian Bischoff Dirk Nowotka
Institute for Formal Methods in Computer Science
Universität Stuttgart, Germany

April 13, 2011

Abstract

An infinte word w avoids a pattern p with the involution θ if there is no substitution for the variables in p and no involution θ such that the resulting word is a factor of w. We investigate the avoidance of patterns with respect to the size of the alphabet. For example, it is shown that the pattern $\alpha\,\theta(\alpha)\,\alpha$ can be avoided over three letters but not two letters, whereas it is well known that $\alpha\,\alpha\,\alpha$ is avoidable over two letters.

1 Introduction

The avoidability of patterns in infinite words is an old area of interest with a first systematic study going back to Thue [4, 5]. This field includes rediscoveries and studies by many authors over the last one hundred years; see for example [2] and [1] for surveys. In this article, we are concerned with a variation of the theme by considering avoidable patterns with involution. An involution θ is a mapping such that θ^2 is the identity. We consider morphic, where $\theta(uv) = \theta(u)\theta(v)$, and antimorphic involutions, where $\theta(uv) = \theta(v)\theta(u)$. The subject of this article draws quite some motivation from applications in biology where the Watson-Crick complement corresponds to an antimorphic involution in our case. Our considerations are more general, however, by considering any alphabet size and also morphic involutions.

2 Preliminaries

Our notation is guided by what is commonly found in literature, see for example the first chapter of [3] as a reference. Let Σ be a finite alphabet of *letters* and Σ^* denote all *finite* and Σ^ω denote all (right-) *infinite* words over Σ. Let ε denote the empty word. Letters are usually denoted by a, b, or c, and words over Σ are usually denoted by u, v, or w in this paper. The i-th letter of a word w is denoted by $w_{[i]}$, that is, $w = w_{[1]}w_{[2]}\cdots w_{[n]}$ if w is finite, and the length n of w is denoted by $|w|$ as usual.

Besides Σ we need another finite set E of symbols. The elements of E are called *variables* and we usually denote them by α, β, or γ. Words in E^* are called *patterns*. For example $\alpha\beta\alpha \in E^*$ is a pattern consisting of the variables α and β in E. We assign to every pattern a *pattern language* over the alphabet Σ. This language contains every word, that can be generated by substituting all variables in the pattern by non-empty words in Σ^*. For example the pattern language of the pattern $\alpha\alpha$ over $\Sigma = \{ a, b \}$ is $\{ aa, bb, aaaa, abab, baba, bbbb, \dots \}$.

We say that a word w *avoids* a pattern, if no factor of w exists, that is in the pattern language. On the other hand, if a factor of w is an element of the pattern language, we say w *contains* the pattern. If for a given pattern e and an alphabet Σ with k elements a word $w \in \Sigma^\omega$ exists that avoids e, then we say that e is *k-avoidable*. Otherwise we call e *k-unavoidable*. We call $k \in \mathbb{N}$ the *avoidance index* $\mathcal{V}(e)$ of a pattern $e \in E^*$, if e is k-avoidable and k is minimal. If no such k exists, we define $\mathcal{V}(e) = \infty$.

Let $f\colon \{ a, b \}^* \to \{ a, b \}^*$ with $a \mapsto ab$ and $b \mapsto ba$. The fixpoint $t = \lim_{k\to\infty} f^k(a)$ exists and is called *Thue–Morse word*. The following result is a classical one.

Theorem 1 ([4, 5])**.** *The Thue–Morse word avoids the patterns $\alpha\alpha\alpha$ and $\alpha\beta\alpha\beta\alpha$.*

3 Patterns with Involution

For introducing patterns with involution, we extend the set of pattern variables E by adding $\theta(\alpha)$ for all variables $\alpha \in E$ and some involution θ. For the rest of the article, we will stick to this definition of E. Given a morphic or antimorphic involution, we build the corresponding pattern language by replacing the variables by non-empty words and, for variables of the form $\theta(\alpha)$, by applying the involution after the substitution.

For example, let θ be the morphic involution with $a \mapsto b$ and $b \mapsto a$ over $\Sigma = \{ a, b \}$, and let the pattern be $\alpha\,\theta(\alpha)$. We get the pattern language $\{ ab, ba, aabb, abba, baab, bbaa, \dots \}$. Every word in $\{ a, b \}^\omega \setminus (a^\omega \cup b^\omega)$ contains the pattern $\alpha\,\theta(\alpha)$ for the morphic involution θ with $a \mapsto b$ and $b \mapsto a$.

Observation 2. *Let θ be a morphic or antimorphic involution and not the identity mapping. Then every pattern, that contains variables of the α and $\theta(\alpha)$, is avoidable.*

Indeed, since θ is not the identity mapping, a letter $a \in \Sigma$ with $\theta(a) \neq a$ exists. Therefore $w = a^\omega$ avoids every pattern that includes variables α and $\theta(\alpha)$.

Because of this observation we do not have to examine, if patterns are avoidable or unavoidable for a given involution. So we now change the point of view. For a given pattern $e \in E^*$, we either look at all morphic or all antimorphic involutions $\Sigma^* \to \Sigma^*$ at the same time. So, we examine, for example, if an infinite word $w \in \Sigma^\omega$ exists, that avoids a pattern e for all morphic involutions.

Definition 3. *Let $e \in E^*$ be a pattern, possibly with variables of the form $\theta(\alpha)$. We call $k \in \mathbb{N}$ the morphic (antimorphic) θ-avoidance index $\mathcal{V}_{\mathrm{m}}^{\theta}(e)$ ($\mathcal{V}_{\mathrm{a}}^{\theta}(e)$) of $e \in E^*$, if an infinite word $w \in \Sigma^\omega$ over Σ with $|\Sigma| = k$ exists, that avoids the pattern e for all morphic (antimorphic) involutions $\Sigma^* \to \Sigma^*$ and k is minimal. If this doesn't hold for any $k \in \mathbb{N}$, we define $\mathcal{V}_{\mathrm{m}}^{\theta}(e) = \infty$ ($\mathcal{V}_{\mathrm{a}}^{\theta}(e) = \infty$).*

We establish the first facts about avoidance of pattern $\alpha\,\theta(\alpha)\,\alpha$.

Lemma 4. *Let Σ be a binary alphabet. Then there is no word $w \in \Sigma^\omega$, that avoids the pattern $\alpha\,\theta(\alpha)\,\alpha$ for all morphic involutions $\theta\colon \Sigma^* \to \Sigma^*$. That is, $\mathcal{V}_{\mathrm{m}}^{\theta}(\alpha\,\theta(\alpha)\,\alpha) > 2$.*

Proof. Let $\Sigma = \{a, b\}$. We try to construct a word $w \in \Sigma^\omega$, that avoids $e = \alpha\,\theta(\alpha)\,\alpha$ for all morphic involutions and bring this to a contradiction. For example, this word must not contain aaa, bbb, aba or bab as a factor. Without loss of generality w begins with a.

Case 1: Assumed the word w begins with ab. Then this prefix must be followed by b, $abb <_{\mathrm{p}} w$. The next letter must be an a, the fifth must be an a too. So we have $abbaa <_{\mathrm{p}} w$. If the following letter is an a, aaa is a factor of w. So the next letter must be the letter b. But for the morphic involution θ with $a \mapsto b$ and $b \mapsto a$ the word $ab\theta(ab)ab$ is a factor of w.

Case 2: The argument for the case $aa \leq_{\mathrm{p}} w$ is analogous to case 1. □

The proof of the following lemma is analogous to the previous one.

Lemma 5. *Let Σ be a binary alphabet. There is no word $w \in \Sigma^\omega$, that avoids the pattern $\alpha\,\theta(\alpha)\,\alpha$ for all antimorphic involutions $\theta\colon \Sigma^* \to \Sigma^*$. That is, $\mathcal{V}_{\mathrm{a}}^{\theta}(\alpha\,\theta(\alpha)\,\alpha) > 2$.*

4 Main Result

In this section, we establish the θ-avoidance indices for the pattern $\alpha\,\theta(\alpha)\,\alpha$ in the morphic and antimorphic case. We start with the morphic case.

Theorem 6. *It holds that $\mathcal{V}_{\mathrm{m}}^{\theta}(\alpha\theta(\alpha)\alpha) = 3$.*

Proof. Let Σ an alphabet with three elements, $\Sigma = \{a, b, c\}$. Let u be the infinitely long Thue–Morse word over the letters a' and b'. Furthermore let $w \in \Sigma$ be the word, that is the outcome of replacing every a' in u by $aacb$ and b' by $accb$. We will show, that w avoids the pattern $\alpha\theta(\alpha)\alpha$ for all morphic involutions. For better readability, we define $x = aacb$ and $y = accb$.

We assume it exists a morphic involution θ and a substitution for α, such that $\alpha\theta(\alpha)\alpha$ is a factor of w. Proof by contradiction. First, we examine the possibilities of replacing the variable α by words $u \in \Sigma^+$ of length $|u| < 7$. The word $u\,\theta(u)\,u$ has a maximal length of 18. Therefore there must exist a morphic involution so that $u\,\theta(u)\,u$ is a factor of a word $w' \in \{x, y\}^6$. Because of Theorem 1, the words xxx, yyy, $xyxyx$ and $yxyxy$ can not be a factor of w'. A computer program can easily check these finite possibilities with the result,

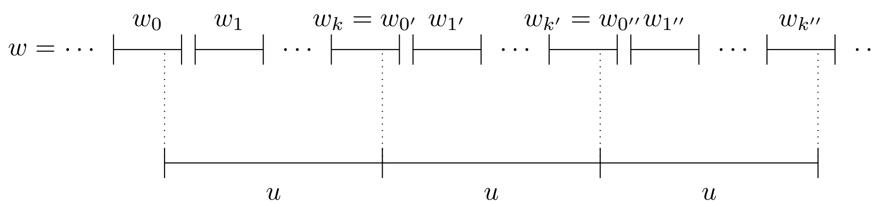

Figure 1: Part of w to illustrate the factor uuu

that no words u and w' exist, which fulfill the conditions. Now we assume α gets replaced by a word $u \in \Sigma^+$ with $|u| \geq 7$. Then, the word u contains $aacb$ or $accb$. Without loss of generality, u contains $aacb$. Therefore, $\theta(u)$ contains the factor $\theta(aac) = \theta(a)\,\theta(a)\,\theta(c)$. In addition $\theta(u)$ and for this reason $\theta(a)\,\theta(a)\,\theta(c)$ is a factor of w. There are only two possibilities for two succeeding identical letters in w. Either these letters are two letters c followed by the letter b, or two letters a are followed by the letter c. This implies, that $u\,\theta(u)\,u$ can only be a factor of w, if θ is the identity mapping. Furthermore this implies $|u| = 4 \cdot k$ for a $k \in \mathbb{N}$. This is visualized in Fig. 1, where $w_i, w_{i'}, w_{i''} \in \{x, y\}$ holds for all $0 \leq i \leq k$. If the word $(w_0)_{[2]}(w_0)_{[3]}(w_0)_{[4]}$ or $(w_0)_{[1]}(w_0)_{[2]}(w_0)_{[3]}(w_0)_{[4]} = w_0$ is a prefix of the first u in Fig. 1, then the following equations apply:

$$\begin{array}{ccccc} w_0 & = & w_{0'} & = & w_{0''} \\ w_1 & = & w_{1'} & = & w_{1''} \\ \vdots & & \vdots & & \vdots \\ w_{k-1} & = & w_{k-1'} & = & w_{k-1''} \end{array}$$

The word $w_0 w_1 \dots w_{k-1}\, w_{0'} w_{1'} \dots w_{k-1'}\, w_{0''} w_{1''} \dots w_{k-1''} = (w_0 w_1 \dots w_{k-1})^3$ is a factor of w. Because of $w_i \in \{x, y\}$ for all $0 \leq i \leq k-1$, this is a contradiction to Lemma 1. On the other hand, if only $(w_0)_{[3]}(w_0)_{[4]}$ or $(w_0)_{[4]}$ is a prefix of u, then $w_0 \neq w_{0'}$ is possible. But in this case $(w_{k''})_{[1]}(w_{k''})_{[2]}$ or $(w_{k''})_{[1]}(w_{k''})_{[2]}(w_{k''})_{[3]}$ is a suffix of the third u. This implies

$$\begin{array}{ccccc} w_1 & = & w_{1'} & = & w_{1''} \\ w_2 & = & w_{2'} & = & w_{2''} \\ \vdots & & \vdots & & \vdots \\ w_k & = & w_{k'} & = & w_{k''} \end{array}$$

and $w_1 w_2 \dots w_k\, w_{1'} w_{2'} \dots w_{k'}\, w_{1''} w_{2''} \dots w_{k''} = (w_1 w_2 \dots w_k)^3$ is a factor of w. Again, this is a contradiction to Lemma 1. The theorem follows with Lemma 4. □

The result of Theorem 6 transfers also to the antimorphic case.

Theorem 7. *It holds that* $\mathcal{V}_{\mathrm{a}}^{\theta}(\alpha\theta(\alpha)\alpha) = 3$.

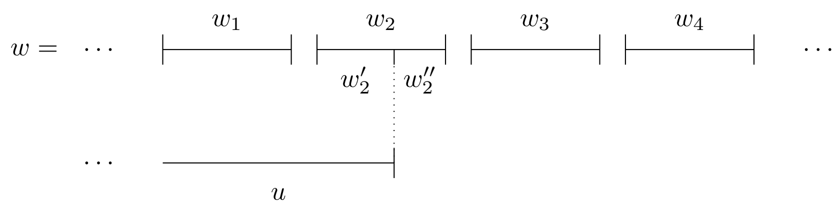

Figure 2: Part of w and the factor u of w

Proof. This proof follows the proof of the previous theorem. Let Σ be an alphabet with three elements, $\Sigma = \{a, b, c\}$. Further, let u be the Thue-Morse word over the letters a' and b'. Let $w \in \Sigma^\omega$ be the word, that we get by replacing a' in u by $aabbc$ and b' by $aaccb$. We will show, that w avoids the pattern $\alpha\,\theta(\alpha), \alpha$ for all antimorphic involutions. For better readability, we define $x = aabbc$ and $y = aaccb$.

We assume that there exists an antimorphic involution and a substitution of α by a word $u \in \Sigma^+$ so, that $u\,\theta(u)\,u$ is a factor of w. First we suppose that $|u| < 9$ holds. The word $u\,\theta(u)\,u$ then has a maximal length of 24 and $u\,\theta(u)\,u$ is factor of a word $w' \in \{x, y\}^6$. The word xxx, yyy, $xyxyx$, and $yxyxy$ must not be a factor of w' because of Lemma 1. A computer program can check these finite possibilities with the result, that no words u and w' exist that fulfill these conditions for an antimorphic involution θ. So ,$|u| \geq 9$ must hold and u contains at least one word x or y completely. We now look at the first u of the factor $u\,\theta(u)\,u$ of w. Let $w_1 w_2' \leq_{\mathrm{s}} u$ with $w_1, w_2 \in \{x, y\}$, $w_2 = w_2' w_2''$ and $|w_2'| < 5$. We get Fig. 2 where $w_3, w_4 \in \{x, y\}$. Without loss of generality, let $w_1 = x = aabbc$. Then $\theta(u)$ and therefore $w_2 w_3 w_4$ contains the word $\theta(aabbc) = \theta(c)\,\theta(b)\,\theta(b)\,\theta(a)\,\theta(a)$ with length 5 as a factor. Hence we look at the following words:

$$\begin{aligned} xx &= aabbc\,aabbc \\ xy &= aabbc\,aaccb \\ yx &= aaccb\,aabbc \\ yy &= aaccb\,aaccb\ . \end{aligned}$$

Only xx contains $\theta(c)\,\theta(b)\,\theta(b)\,\theta(a)\,\theta(a)$ for the antimorphic involution θ with $a \mapsto b$, $b \mapsto a$, and $c \mapsto c$. Because of $w_1 = x$, the equation $w_2 w_3 = xx$ is a contradiction to Lemma 1. The case $w_2 w_3 w_4 = yxx$ remains. Now there are five possibilities for the position of u, see Fig. 3. It is easy to check, that in all five cases $\theta(u) \leq_{\mathrm{p}} w_2'' w_3 w_4$ respectively $w_2'' w_3 w_4 \leq_{\mathrm{p}} \theta(u)$ doesn't hold. So our assumption, that there exists an antimorphic involution θ and a word $u \in \Sigma^+$ with $u\,\theta(u)\,u$ is a factor of w, was wrong. The theorem follows with Lemma 5. ☐

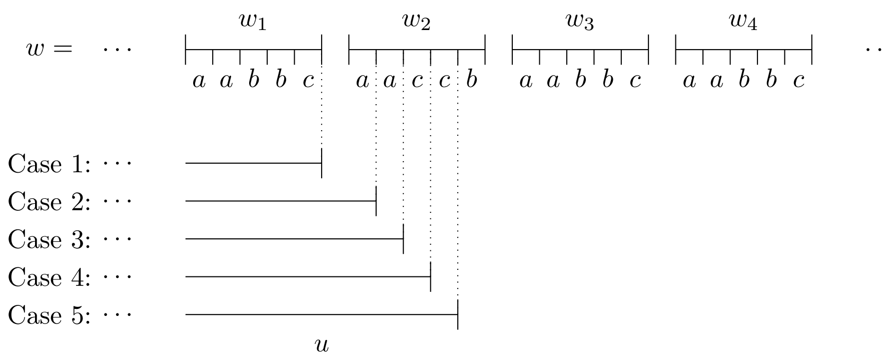

Figure 3: Illustration of possible positions of the factor u of w

5 Complementary Patterns

In this section, patterns similar to $\alpha\,\theta(\alpha)\,\theta$ are considered.

For the next lemma we need a further definition. Let $e \in E^*$ be a pattern consisting of variables of the form α and $\theta(\alpha)$ and e' be the pattern that we get, when all variables α and $\theta(\alpha)$ in e are switched. We call $e' \in E$ the *θ-complementary* pattern of e. For example the θ-complementary pattern of $\alpha\,\alpha\,\theta(\alpha)\,\beta$ is $\theta(\alpha)\,\theta(\alpha)\,\alpha\,\theta(\beta)$. For this definition it doesn't matter if morphic or antimorphic involutions are examined.

Lemma 8. *Let $e \in E^*$ be a pattern and $e' \in E$ be the θ-complementary pattern of e. Then $\mathcal{V}_{\mathrm{a}}^{\theta}(e) = \mathcal{V}_{\mathrm{a}}^{\theta}(e')$ and $\mathcal{V}_{\mathrm{m}}^{\theta}(e) = \mathcal{V}_{\mathrm{m}}^{\theta}(e')$.*

Proof. First of all we show $\mathcal{V}_{\mathrm{m}}^{\theta}(e) = \mathcal{V}_{\mathrm{m}}^{\theta}(e')$. For better readability, we replace the variable α in the pattern e' by α' and $\theta(\alpha)$ by $\theta(\alpha')$. We assume a word $w \in \Sigma^\omega$ contains the pattern e for a morphic involution and a substitution of α by $u \in \Sigma^+$. Then w contains the pattern e' for the same morphic involution by substituting α' by $\theta(u)$. Symmetry reasons imply:

It exists a morphic involution θ so, that w contains the pattern e.
$\Leftrightarrow$ It exists a morphic involution θ' so, that w contains the pattern e'.

By negation we get:

The word $w \in \Sigma^\omega$ avoids the pattern e.
$\Leftrightarrow$ The word $w \in \Sigma^\omega$ avoids the pattern e'.

The equation $\mathcal{V}_{\mathrm{m}}^{\theta}(e) = \mathcal{V}_{\mathrm{m}}^{\theta}(e')$ follows. The proof of $\mathcal{V}_{\mathrm{a}}^{\theta}(e) = \mathcal{V}_{\mathrm{a}}^{\theta}(e')$ is identical. □

Note the following θ-free patterns; see [1].

Observation 9. *The patterns* $\alpha\alpha$, $\alpha\alpha\beta$, $\beta\alpha\alpha$, $\alpha\alpha\beta\alpha$, $\alpha\beta\beta\alpha$, $\alpha\alpha\beta\beta$, $\alpha\beta\alpha\beta$, $\alpha\alpha\beta\alpha\alpha$, *and* $\alpha\alpha\beta\alpha\beta$ *are* 2*-unavoidable and* 3*-avoidable.*

Lemma 10. *Let* $e \in E^*$ *be a pattern, that contains the variables* α *and* $\theta(\alpha)$*. Further,* e *contains no other variable of the form* $\theta(\gamma)$*. Let* e' *be the pattern when all occurrences of* $\theta(\alpha)$ *in* e *are replaced by* α*. The pattern* e'' *obtained when all occurrences of* $\theta(\alpha)$ *in* e *are replaced by a new variable* β*.*

Then $\mathcal{V}(e') \leq \mathcal{V}^{\theta}_{\mathrm{m}}(e) \leq \mathcal{V}(e'')$ *and* $\mathcal{V}^{\theta}_{\mathrm{a}}(e) \leq \mathcal{V}(e'')$*.*

Proof. The relation $\mathcal{V}(e') \leq \mathcal{V}^{\theta}_{\mathrm{m}}(e)$ holds, since the morphic θ-avoidance index considers all morphic involutions, including the identity mapping. Now say $\mathcal{V}(e'') = k$, i.e., a word $w \in \Sigma^{\omega}$ exists, that avoids the pattern e''.Then this word also avoids the pattern e for all morphic and antimorphic involutions. Therefore the relations $\mathcal{V}^{\theta}_{\mathrm{m}}(e) \leq \mathcal{V}(e'')$ and $\mathcal{V}^{\theta}_{\mathrm{a}}(e) \leq \mathcal{V}(e'')$ hold. □

Lemma 11. *It holds that* $\mathcal{V}^{\theta}_{\mathrm{a}}(\alpha\,\alpha\,\theta(\alpha)) = \mathcal{V}^{\theta}_{\mathrm{m}}(\alpha\,\alpha\,\theta(\alpha)) = 3$*.*

Proof. According to Observation 9 the equation $\mathcal{V}(\alpha, \alpha\,\beta) = 3$ holds. Lemma 10 implies $\mathcal{V}^{\theta}_{\mathrm{a}}(\alpha\,\alpha\,\theta(\alpha))\,, \mathcal{V}^{\theta}_{\mathrm{m}}(\alpha\,\alpha\,\theta(\alpha)) \leq 3$. We show by contradiction, that it holds that $\mathcal{V}^{\theta}_{\mathrm{a}}(\alpha\,\alpha\,\theta(\alpha)) \neq 2$. The proof for the relation $\mathcal{V}^{\theta}_{\mathrm{m}}(\alpha\,\alpha\,\theta(\alpha)) \neq 2$ is analogous. Assuming a word $w \in \Sigma^{\omega}$ with $\Sigma = \{\,a, b\,\}$ exists that avoids the pattern $\alpha\,\alpha\,\theta(\alpha)$ for all antimorphic involutions. Then w contains neither aa nor bb as a factor. Without loss of generality w begins with the letter a. It follows that $w = (ab)^{\omega}$. But $w = (ab)^{\omega}$ contains the pattern $\alpha\,\alpha\,\theta(\alpha)$ for $\alpha = ab$ and the antimorphic involution defined by $a \mapsto b$ and $b \mapsto a$. This is a contradiction to our assumption. Therefore $\mathcal{V}^{\theta}_{\mathrm{a}}(\alpha\,\alpha\,\theta(\alpha)) \neq 2$ holds and analogously $\mathcal{V}^{\theta}_{\mathrm{m}}(\alpha\,\alpha\,\theta(\alpha)) \neq 2$. We get $\mathcal{V}^{\theta}_{\mathrm{a}}(\alpha\,\alpha\,\theta(\alpha)) = \mathcal{V}^{\theta}_{\mathrm{m}}(\alpha\,\alpha\,\theta(\alpha)) = 3$. □

Lemma 12. *It holds that* $\mathcal{V}^{\theta}_{\mathrm{a}}(\theta(\alpha)\,\alpha\,\alpha) = \mathcal{V}^{\theta}_{\mathrm{m}}(\theta(\alpha)\,\alpha\,\alpha) = 3$*.*

Proof. The proof is analogous to the proof of Lemma 11. □

Corollary 13.

1. $\mathcal{V}^{\theta}_{\mathrm{m}}(\theta(\alpha)\,\alpha\,\theta(\alpha)) = \mathcal{V}^{\theta}_{\mathrm{a}}(\theta(\alpha)\,\alpha\,\theta(\alpha)) = 3$ *by Theorem 6 and 7.*
2. $\mathcal{V}^{\theta}_{\mathrm{m}}(\theta(\alpha)\,\theta(\alpha)\,\alpha) = \mathcal{V}^{\theta}_{\mathrm{a}}(\theta(\alpha)\,\theta(\alpha)\,\alpha) = 3$ *by Lemma 11.*
3. $\mathcal{V}^{\theta}_{\mathrm{m}}(\alpha\,\theta(\alpha)\,\theta(\alpha)) = \mathcal{V}^{\theta}_{\mathrm{a}}(\alpha\,\theta(\alpha)\,\theta(\alpha)) = 3$ *by Lemma 12.*

References

[1] J. Cassaigne. *Unavoidable Patterns*, chapter 3, pages 111–134. In [3], 2002.

[2] J. Currie. Pattern avoidance: themes and variations. *Theoret. Comput. Sci.*, 339(1):7–18, 2005.

[3] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, Cambridge, UK, 2002.

[4] A. Thue. Über unendliche Zeichenreihen. *Norske Vid. Skrifter I. Mat.-Nat. Kl., Christiania*, 7:1–22, 1906.

[5] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske Vid. Skrifter I. Mat.-Nat. Kl., Christiania*, 1:1–67, 1912.