Combinatorics/Computer Science

Invertible substitutions on a three-letter alphabet

Substitutions inversibles sur un alphabet de trois lettres

Bo Tan, Zhi-Xiong Wen, Yiping Zhang

Department of Mathematics, Wuhan University, 430072 Wuhan, Hubei, PR China

Received 29 October 2002; accepted 5 November 2002

Presented by Jean-Pierre Kahane

Abstract

We study the structure of invertible substitutions on a three-letter alphabet. We show that there exists a finite set \( S \) of invertible substitutions such that any invertible substitution can be written as \( I_w \circ \sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_k \), where \( I_w \) is the inner automorphism associated with \( w \), and \( \sigma_j \in S \) for \( 1 \leq j \leq k \). As a consequence, \( M \) is the matrix of an invertible substitution if and only if it is a finite product of non-negative elementary matrices.

Résumé

Nous étudions la structure des substitutions inversibles sur un alphabet à trois lettres. Nous prouvons qu’il existe un ensemble fini \( S \) de substitutions inversibles tel que toute substitution inversible puisse être écrite comme \( I_w \circ \sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_k \), où \( I_w \) est l’automorphisme intérieur associé à \( w \) et où \( \sigma_j \in S \) pour \( 1 \leq j \leq k \). Comme conséquence, \( M \) est la matrice d’une substitution inversible si et seulement si elle est un produit fini de matrices élémentaires non-négatives.

Version française abrégée

Soit \( A = \{a, b, c\} \) un alphabet de trois lettres. On désigne par \( A^* \) et \( \Gamma_A \) les monoïde et groupe libres engendrés par \( A \). Une substitution \( \sigma \) sur \( A \) (c’est-à-dire un endomorphisme de \( A^* \)) est dite inversible si elle se prolonge en un automorphisme de \( \Gamma_A \). On peut identifier \( \sigma \) au triplet de mots \( (\sigma(a), \sigma(b), \sigma(c)) \). On pose \( \pi_1 = (b, a, c) \), \( \pi_2 = (c, b, a) \), \( \phi_l = (ba, b, c) \), \( \phi_r = (ab, b, c) \). Le monoïde des substitutions inversibles sur \( A \) sera noté \( IS(A^*) \).

Si \( w \in \Gamma_A \), on désigne par \( I_w \) l’automorphisme intérieur de \( \Gamma_A \) associé à \( w \) : \( I_w(u) := wuw^{-1} \).

La structure du monoïde des substitutions inversibles sur un alphabet de deux lettres est connue : il est engendré par une permutation et deux substitutions de Fibonacci [12]. Quand l’alphabet a plus de deux lettres, la situation...
est beaucoup plus compliquée. Dans [15], il est prouvé que le monoïde IS(A∗) n’est pas de type fini, mais sa structure restait jusqu’à présent inconnue.

Dans cette Note, nous prouvons les théorèmes suivants.

**Théorème 1.** Soit σ est une substitution inversible. Il existe w ∈ A∗ ou w−1 ∈ A∗ et σ1, . . . , σk ∈ {π1, π2, φl, φr} tels que

1. σ = Iw ◦ σ1 ◦ . . . ◦ σk.
2. De plus, on peut choisir pour w (ou w−1) un suffixe (ou un préfixe) commun aux mots σ(a), σ(b) et σ(c).

**Théorème 2.** Une matrice 3 × 3 à coefficients entiers positifs ou nuls est la matrice d’une substitution inversible si et seulement si c’est un produit fini de matrices élémentaires à coefficients positifs.

1. Introduction

   Le study of substitutions (the endomorphisms of free monoids of finite type) plays an important role in finite automata, symbolic dynamics, and fractal geometry [1,2,4]. It has various applications to quasicrystals, computational complexity, information theory. In addition, substitution is also a fundamental object studied in combinatorial group theory [6–8].

   Plenty of results have been obtained for substitutions over a two-letter alphabet [3,11,13]. The notion of invertible substitution appears in [10]: these are the substitutions which extend as automorphisms of the corresponding free group. Since then, they have been studied by many authors (for instance, see [3,5]). The invertible substitutions over a two-letter alphabet form a monoid whose structure in known [12]: this monoid can be generated by a permutation and two so-called Fibonacci substitutions. This result has had many applications, namely to the study of local isomorphisms of fixed points of substitutions [12,14] and to the study of trace maps [10].

   When the alphabet has more than two letters, the situation is much more complicated. In [15], it was shown, by enumerating infinitely many so-called indecomposable substitutions, that the monoid of invertible substitutions over three letters (which will be denoted by IS(A∗)) is not finitely generated. But, up to now the structure of IS(A∗) remained unknown. Here, we elucidate this structure. We show that if σ is an invertible substitution over a three-letter alphabet, there exists a word w such that Iw ◦ σ or Iw−1 ◦ σ is the composition of finitely many Fibonacci substitutions and permutations (Theorem 3.1). As a consequence, the matrix of an invertible substitution is positively decomposable (Theorem 3.2).

2. Preliminaries and notations

   Let us first recall some basic definitions and notations in the theory of substitutions (see [1,8] for a general theory).

   Set A = {a, b, c} and A = {a, b, c, a−1, b−1, c−1}. Let A∗ (resp. ΓA) be the free monoid (resp. the free group) generated by A (the unit element is the empty word ε). The elements of A∗ will be called “positive words” or simply “words” and those of ΓA “signed” or “mixed” words. The inverse of a positive word will be said to be “negative”.

   Let w ∈ ΓA: w = x1 · · · xk with xi ∈ A (i = 1, 2, . . . , k). If xi,xi+1 ̸= ε (for i = 1, . . . , k − 1), we say that x1 · · · xk is in the reduced form and that the length of w is k. This length will be denoted by |w|. Let w, wj (j = 1, . . . , k) ∈ ΓA. If w = w1w2 · · · wk satisfies |w| = |w1| + · · · + |wk|, we say that w1w2 · · · wk is a
reduced expression of \( w \). Then, we say that \( w_1 \) is a prefix of \( w \) and that \( w_k \) is a suffix of \( w \), and we write \( w_1 < w \) and \( w_k > w \).

A substitution \( \sigma \) over \( A \) is a morphism \( \sigma \) of \( A^* \). Such a morphism extends in a natural way to an endomorphism of \( \Gamma_A \); If this extension is an automorphism of \( \Gamma_A \), the substitution \( \sigma \) is said to be invertible. The set of substitutions (resp. invertible substitutions) is denoted by \( S(A^*) \) (resp. by \( IS(A^*) \)).

We often identify an endomorphism \( \sigma \) of \( \Gamma_A \) with the triple \((\sigma(a), \sigma(b), \sigma(c))\) of (maybe mixed) words. We define the length of \( \sigma \) to be \( |\sigma| = |\sigma(a)| + |\sigma(b)| + |\sigma(c)| \).

If \( U \) is a subset of a monoid, \( (U) \) stands for the sub-monoid (not the sub-group, even when dealing inside a group) generated by \( U \). The use of the same notation for different groups and monoids will not generate any confusion.

We shall use the following basic invertible substitutions and automorphisms.

**Permutations**: Let \( \mathcal{P} = (\pi_1, \pi_2) \), where \( \pi_1 = (b, a, c) \) and \( \pi_2 = (c, b, a) \). Note that \( \mathcal{P} \) is a subgroup isomorphic to the symmetric group on \( A \).

**Fibonacci type**: Set \( \phi = (b, a, c), \phi' = (ab, b, c) \). Let \( \mathcal{L} = \{ \pi \circ \phi \circ \pi' \colon \pi, \pi' \in \mathcal{P} \}, \mathcal{R} = \{ \pi \circ \phi_l \circ \pi' \colon \pi, \pi' \in \mathcal{P} \}, \) and \( \mathcal{F} = \mathcal{L} \cup \mathcal{R} \). The elements in \( \mathcal{F} \) are called substitutions of Fibonacci type or simply Fibonacci substitutions.

**Simple substitutions**: Set \( S = (\pi_1, \pi_2, \phi_l, \phi_r) = (\mathcal{P}, \mathcal{F}, \mathcal{I}) \subset IS(A^*) \). The elements in \( S \) will be called simple substitutions.

**Involutions**: \( I_1 = (a^{-1}, b, c), I_2 = (a, b^{-1}, c), I_3 = (a, b, c^{-1}) \), \( \mathcal{I} = \{ I_i \colon 1 \leq i \leq 3 \} \).

According to Nielsen’s theory [8], \( \text{Aut}(\Gamma_A) = (\mathcal{P}, \mathcal{F}, \mathcal{I}) = (\pi_1, \pi_2, \phi_l, \phi_r, I_1) \).

**Definition 2.1.** An invertible substitution \( \sigma \) is said to be trivial if \( \sigma \in \mathcal{P} \). If there exist non-trivial invertible substitutions \( \sigma_1 \) and \( \sigma_2 \) such that \( \sigma = \sigma_1 \circ \sigma_2 \), we say that \( \sigma \) is decomposable.

The invertible substitutions which are not in \( \mathcal{P} \cup \mathcal{F} \) and not decomposable will be called non-simple indecomposable. As a matter of fact, there exist infinitely many such substitutions, hence \( IS(A^*) \) is not finitely generated [15].

**Inner automorphisms**: Let \( z \in \Gamma_A \). \( I_z \in \text{Aut}(\Gamma_A) \) is defined as follows: \( I_z(w) = zwz^{-1} \) (\( w \in \Gamma_A \)). That is \( I_z = (za^{-1}, zb^{-1}, zc^{-1}) \). We have \( I_z \circ (w_1, w_2, w_3) = (zw_1z^{-1}, zw_2z^{-1}, zw_3z^{-1}) \) and \( \sigma \circ I_z = I_{\sigma(z)} \circ \sigma \). Notice also that \( I_e = I \).

If \( w \in \Gamma_A \), one denotes by \( |w|_a \) (resp. \( |w|_b \) or \( |w|_c \)) the number (the algebraic sum of exponents) of appearances of \( a \) (resp. \( b \) or \( c \)) in \( w \) (e.g., \( |a^{-1}ba|_a = 0 \)). Let \( \sigma \in S(A^*) \). One associates a matrix \( M_\sigma \) with \( \sigma: M_\sigma = (\langle \sigma(\beta) \rangle)_{\alpha, \beta \in A} \). Let \( \sigma, \tau \in S(A^*) \). One has \( M_{\sigma \tau} = M_\sigma M_\tau \). If \( \sigma \in IS(A^*) \), then \( \text{det}(M_\sigma) = \pm 1 \).

The above definitions and equalities can be extended to the case when \( \sigma \) and \( \tau \) are endomorphisms of \( \Gamma_A \).

### 3. Main theorems

The following decomposition theorem characterizes the structure of \( IS(A^*) \): any invertible substitution is a simple one up to an inner automorphism.

**Theorem 3.1.** Let \( \sigma \in IS(A^*) \). There exists \( w \in A^* \) or \( w^{-1} \in A^* \) such that

\[
(1) \quad I_w \circ \sigma \text{ is a simple substitution. That is, there exist } \sigma_1, \ldots, \sigma_k \in \{ \pi_1, \pi_2, \phi_l, \phi_r \} \text{ such that } \sigma = I_{w^{-1}} \circ \sigma_1 \cdots \sigma_k.
\]
Furthermore, we can take \( w \) (or \( w^{-1} \)) to be a common suffix (or prefix) of \( \sigma(a) \), \( \sigma(b) \), and \( \sigma(c) \).

**Remark 3.1.** We point out that the decomposition theorem for the case of two-letter alphabet [12] is an easy consequence of Theorem 3.1.

As consequence we have:

**Theorem 3.2.** A \( 3 \times 3 \)-matrix with non-negative integer coefficients is the matrix of some invertible substitution if and only if it is a finite product of non-negative elementary matrices.

### 4. Proofs

The symbol “+” (resp. “−”) will represent various non-empty positive words (resp. non-empty negative words). We also use the symbols like “++”, “−−” to represent various types of mixed words. As an example, \( u = +−+ \) means \( u = u_1 u_2^{-1} u_3 \), where \( u_j \in A^* \) and \( u_1 u_2^{-1} u_3 \) is a reduced expression. The meanings of “+”, “−”, “++”, “−−”, “+−+” etc. are now clear.

The following lemma is a simple version of Nielsen’s cancellation procedure that we use in the proofs. For the details, we refer the reader to [8,9].

**Lemma 4.1.** Let \( \sigma = (w_1, w_2, w_3) \in \text{Aut}(\Gamma_A) \). There exist \( k \geq 0 \) and \( \tau_1, \ldots, \tau_k \in \{ \pi_1, \pi_2, \phi_l, \phi_r, \iota \} \) such that, if one sets \( \sigma_0 = \sigma \) and \( \sigma_i = \sigma_{i-1} \circ \tau_i \) (for \( i = 1, \ldots, k \)), one has \( |\sigma_i| \leq |\sigma_{i-1}| \) (\( i = 1, \ldots, k \)) and \( \sigma_k \) is the identity.

**Definition 4.1.** We say that a non-trivial substitution \( \sigma = (w_1, w_2, w_3) \) is mixed if it satisfies \( w_i w_j^{-1} w_k = +−+ \) for all \((i,j,k)\) such that \( i \neq j \) and \( j \neq k \).

**Lemma 4.2.** Any mixed substitution is non-invertible.

**Corollary 4.1.** Let \( u \in \Gamma_A \) and \( x, y \in A^* \) non-empty. Suppose that \( u = x y \) and \( |u| = |x| + |y| \). Then we have either \( u \in A^* \) or \( u^{-1} \in A^* \).

When \( w = \alpha_1 \alpha_2 \cdots \alpha_{|w|} \in \Gamma_A \) (\( \alpha_i \in \bar{A} \)), we shall use the following notations:

\[
\begin{align*}
 h_i(w) &= \alpha_i \quad (i = 1, 2, \ldots, |w|), \\
 h_\infty(w) &= h_{|w|}(w).
\end{align*}
\]

The following lemmas study the “cancellation properties” between special words.

**Lemma 4.4.** Let \( u, v, x \) and \( y \) be non-empty words satisfying

\[
\begin{align*}
 h_1(u) \neq h_1(y) \quad \text{and} \quad h_\infty(x) \neq h_\infty(v), & \quad (4.1) \\
 x \ (\text{resp. } v) \text{ is not a prefix of } v \ (\text{resp. } x), & \quad (4.2) \\
 y \ (\text{resp. } u) \text{ is not a suffix of } u \ (\text{resp. } y). & \quad (4.3)
\end{align*}
\]

Set \( w_1 = ux, w_2 = uv, w_3 = yv \) and consider the following mixed word:

\[
w = w_1^{-\varepsilon_1} w_2^{-\varepsilon_2} \cdots w_{|w|}^{(-1)^{\varepsilon_k}},
\]

(4.4)
where $k \geq 0$, $\varepsilon \in \{+1, -1\}$, $i_m \in \{1, 2, 3\}$ ($m = 0, 1, \ldots, k$) and $i_m \neq i_{m+1}$ ($m = 0, 1, \ldots, k - 1$). Then we have

$$h_1(w) = h_1(w_{i_0}^{\varepsilon}), \ h_\infty(w) = h_\infty(w_{i_k}^{(-1)^{i_k} \varepsilon}),$$

$$|w| \geq 2. \tag{4.5}$$

The above lemma will be used to study substitutions of the form $(ux, uv, yv)$. The following lemma is the corresponding version for $(uxv, uv, y)$. 

**Lemma 4.5.** Let $u$, $v$ and $y$ be non-empty words such that $h_1(u) \neq h_1(y)$, $h_\infty(v) \neq h_\infty(y)$, $v$ is not a prefix of $xv$ and $u$ is not a suffix of $ux$.

Set $w_1 = uxv$, $w_2 = uv$, $w_3 = y$ and consider the mixed words of the form $(4.4)$. Then

(i) $(4.5)$ holds;
(ii) If $|y| > 1$, $(4.6)$ holds;
(iii) If $|y| = 1$, $(4.6)$ holds except when $k = 0$, and $w_{i_0} = w_3$.

**Lemma 4.6.** Under the notations and conditions of Lemma 4.4 (resp. Lemma 4.5), the substitution $\sigma = (w_1, w_2, w_3)$ is not invertible.

**Proposition 4.1.** Suppose that $\sigma = (w_1, w_2, w_3)$ is a non-simple indecomposable substitution. Then we have either $h_1(w_1) = h_1(w_2) = h_1(w_3)$ or $h_\infty(w_1) = h_\infty(w_2) = h_\infty(w_3)$. In other words, $w_1$, $w_2$, and $w_3$ must have a common non-empty prefix or suffix.

**Lemma 4.7.** Let $z, u, v, x, y \in A^*$. Assume that $u, v, x, y$ satisfy $(4.1)$, $(4.2)$ and $(4.3)$. Set $w_1 = ux$, $w_2 = uv$, $w_3 = yv$, then both substitutions $(zw_1, zw_2, zw_3)$ and $(w_1z, w_2z, w_3z)$ are not invertible.

**Remark 4.1.** When $x, y, u, v, w_1, w_2, w_3$ are given as in Lemma 4.5, conclusions similar to those of the above lemma hold.

**Lemma 4.8.** Suppose that $\sigma = (zu_1, zu_2, zu_3)$ (resp. $\sigma = (u_1z, u_2z, u_3z)$) is a non-simple indecomposable substitution, where $z$ is a non-empty word. Suppose that $u_1, u_2, u_3$ have no common prefix and no common suffix.

Then there exist a non-trivial invertible substitution $\sigma'$, a permutation $\pi \in P$ and a Fibonacci $f \in F$, such that $I_{z^{-1}} \circ \sigma = \sigma' \circ f \circ \pi$ (resp. $I_z \circ \sigma = \sigma' \circ f \circ \pi$).

**Lemma 4.9.** Let $\sigma$ be a non-simple indecomposable substitution. Then there exists $z \in \Gamma_A$ such that $I_z \circ \sigma$ is decomposable. That is, there exist non-trivial invertible substitutions $\sigma_1$ and $\sigma_2$ such that $I_z \circ \sigma = \sigma_1 \circ \sigma_2$.

Furthermore, $|\sigma_i| < |\sigma|$ ($i = 1, 2$).

Now we are ready to prove our theorems.

**Proof of Theorem 3.1.** First we prove Theorem 3.1(1). Let $\sigma$ be an invertible substitution. Then there exist $k \geq 1$ and indecomposable substitutions $\sigma_1, \ldots, \sigma_k$ such that $\sigma = \sigma_1 \circ \cdots \circ \sigma_k$.

If some $\sigma_i$ is non-simple, then by Lemma 4.9, there exist $z \in \Gamma_A$, invertible substitutions $\sigma_i'$ and $\sigma_i''$ such that $\sigma_i = I_{z^{-1}} \circ \sigma_i' \circ \sigma_i''$ and that

$$|\sigma_i'| < |\sigma_i|, \quad |\sigma_i''| < |\sigma_i|. \tag{4.7}$$

Then we repeat such decomposition for $\sigma_i'$ (resp. $\sigma_i''$) and so on. By (4.7), such decomposition will terminate after a finite of steps. Finally every factor will be a simple substitution. Hence we can write $\sigma = I_{w_1} \circ \tau_1 \circ I_{w_2} \circ \tau_2 \circ$
\[ \cdots \circ I_{w_n} \circ \tau_n, \text{ where } \tau_i (i = 1, \ldots, n) \text{ is a simple substitution and } w_i \in \Gamma_A \text{ (put } w_1 = \varepsilon \text{ if necessary). Thus we may write } \sigma = I_{w_1} \circ \tau_1 \circ \tau_2 \circ \cdots \circ \tau_n, \text{ where } w \in \Gamma_A. \]

Let \( \tau = \tau_1 \circ \tau_2 \circ \cdots \circ \tau_n \). It is clear that \( \tau \) is a simple substitution.

Finally, since \( I_{w^{-1}} \circ \sigma = \tau \), Lemma 4.3 implies that \( w \in A^* \) or \( w^{-1} \in A^* \). Theorem 3.1(1) is thus proved.

For Theorem 3.1(2), denoting \( |\sigma|_{\text{min}} = \min\{|\sigma(a)|, |\sigma(b)|, |\sigma(c)|\} \), we only need to prove the following equivalent statement.

**Claim.** We can choose \( w \) in (1) such that \( |w| \leq |\sigma|_{\text{min}} \).

Let us prove the claim by induction on \( k = |\sigma| \).

If \( k \leq 4 \) the claim can be verified simply by enumerating all cases.

Suppose that the claim is true for \( k \leq n \) and that \( |\sigma| = n + 1 \). By the conclusion (1) of the theorem, there exists \( w \in A^* \) or \( w^{-1} \in A^* \) such that \( I_w \circ \sigma \) is simple. To be specific, we can suppose that \( w^{-1} \in A^* \), \( \sigma = (w_1, w_2, w_3) \) and that \( |w_1| = |\sigma|_{\text{min}} \).

If \( |w| \leq |\sigma|_{\text{min}} \), nothing needs to be proven. Suppose \( |w| > |\sigma|_{\text{min}} = |w_1| \). It is easy to see that \( w_1 \) is a common prefix of \( w_2, w_3 \). Let then \( w_2 = w_1 \w_2' \). We have \( (w_1, w_2, w_3) = (w_1, w_2', w_2) \circ (a, ab, c) \), that is, \( \sigma = \sigma' \circ f \), where \( \sigma' = (w_1, w_2', w_2) \) is obviously an invertible substitution and \( f = (a, ab, c) \) is a Fibonacci substitution.

It is trivial that \( |\sigma'| < |\sigma'| = n + 1 \) and that \( |\sigma'|_{\text{min}} \leq |\sigma|_{\text{min}} \), hence by the hypothesis of induction we have the following fact:

There exists \( z \in A^* \) (or \( z^{-1} \in A^* \)) such that \( |z| \leq |\sigma'|_{\text{min}} \) and that \( g := I_z \circ \sigma' \) is simple.

Then it follows that \( I_z \circ \sigma = I_z \circ \sigma' \circ f = g \circ f \).

Since \( |z| \leq |\sigma'|_{\text{min}} \leq |\sigma|_{\text{min}} \) and \( g \circ f \) is (by definition) simple, the conclusion is proved.

**Acknowledgement**

The authors would like to thank Professors S. Ito, J. Peyrière, and Z.-Y. Wen for helpful discussions.

**References**