On fairness of $D0L$ systems

Anton Černý

Department of Mathematics and Computer Science, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait

Available online 14 April 2007

Abstract

A word is called fair if it contains, for each pair of distinct symbols $a$, $b$, the same number of occurrences of the scattered subword $ab$ as of $ba$. We prove that if the first $k + 1$ words in the sequence generated by a $D0L$ system over a $k$-letter alphabet are fair then all words in the sequence are fair.

Keywords: Parikh mapping; $D0L$ system; Fair word

1. Introduction

Palindromes—the words identical to their mirror image—have been studied in various aspects. We would like to focus here on one particular property of palindromes. The apparent symmetry implies, that whatever pair of distinct symbols $a$, $b$ we choose, there is the same number of those its occurrences where $a$ precedes $b$ as those where $b$ precedes $a$. But palindromes are not the only words with this property. For example, the word $ab^3a^2b$ contains 6 different occurrences of the pair $a$, $b$ where $a$ precedes $b$, and the same number of occurrences with $b$ preceding $a$. Such words are called fair, since they do not prefer any of the two symbols in the order of appearance [1]. For an alphabet of size at least 2, the language of all palindromes is well known to be context-free, but not regular. The language of all fair words over an alphabet of size at least 2 is context-sensitive, but not context-free (the argument stated in [9] for a binary alphabet is valid for alphabets of any larger size). It is natural to look for a simple tool to generate fair words of arbitrary size. The $D0L$ system (morphism iteration) seems to be one of the natural candidates. Our result shown here documents, that we can guarantee fairness of the words generated by a $D0L$ system. We prove, that if the first $k + 1$ words in a $D0L$ sequence are fair, then all words in the sequence are fair. This is a generalization of the result from [8].

Recently, Parikh matrices [2–6], a generalization of the Parikh mapping [7] proved to be a very useful tool for investigation of the subword structure of words. In our proof, we use $p$-matrices [1] being a different generalization of the Parikh mapping, and balance-matrices derived from them.
2. Basic terms

In the following, $\Sigma = \{a_1, a_2, \ldots, a_k\}$, $A = \{b_1, \ldots, b_m\}$ denote two ordered alphabets, $k, m \geq 1$ and $h : \Sigma^* \rightarrow A^*$ a morphism. All vectors used are row vectors.

We denote by $|w|_u$ the number of occurrences of the (scattered—not necessarily being a contiguous factor) subword $u$ in the word $w$. For example, $|bab}_{ab}ba|_{ab} = 4$, $|bab}_{ba}ba|_{bab} = 6$. A word $w$ is fair, if for any two distinct symbols $a, b$ occurring in $w$, $|w|_{ab} = |w|_{ba}$.

A DOL system is a triple $G = (\Gamma, \psi, z)$ where $\Gamma$ is an alphabet, $\psi : \Gamma^* \rightarrow \Gamma^*$ is a morphism and $z \in \Gamma^*$ is the axiom of $G$. The DOL sequence $S(G)$ defined by $G$ is the infinite sequence $z, \psi(z), \psi^2(z), \psi^3(z), \ldots$. The sequence $S(G)$ is fair if each of its elements is a fair word.

The Parikh vector of a word $w \in \Sigma^*$ is the vector $p_w^\Sigma = (|w|_{a_1}, \ldots, |w|_{a_k})$.

We consider the set $\mathcal{F}_k$ of all $k \times k$ matrices on the set $\mathbb{Z}$ of all integers. The zero matrix from $\mathcal{F}_k$ is denoted as $0^{(k)}$. For a matrix $X$ and $1 \leq i, j \leq k$, we denote as $X_{ij}$ the element of $X$ at position $i, j$. Similarly, $x_i$ denotes the $i$th element of a vector $x$. The following operation $\circ$ is defined on $\mathcal{F}_k$. For $A, B \in \mathcal{F}_k$,

$$(A \circ B)_{i,j} = \begin{cases} A_{i,i} + B_{i,i} & \text{if } i = j, \\ A_{i,j} + B_{i,j} + A_{i,i}B_{j,j} & \text{if } i \neq j. \end{cases}$$

The $p$-matrix of a word $w \in \Sigma^*$ is the matrix $P_w^\Sigma \in \mathcal{F}_k$ where, for $1 \leq i, j \leq k$, $(P_w^\Sigma)_{i,j} = |w|_{a_i}$ if $i = j$ and $(P_w^\Sigma)_{i,j} = |w|_{a_ia_j}$ otherwise. From [1] we have ($x^R$ denotes the mirror image of $x$):

**Proposition 1.** 1. $(\mathcal{F}_k, \circ)$ is a (non-commutative, for $k \geq 2$) group. For $A, B \in \mathcal{F}_k$, $(A \circ B)^T = B^T \circ A^T$.

2. For $x, y \in \Sigma^*$, $P_x^\Sigma \circ P_y^\Sigma = P_{xy}^\Sigma$ and $P_{X^R}^\Sigma = (P_X^\Sigma)^T$.

Let $u, v$ be two $k$-dimensional ($k \geq 1$) integer vectors. We define the matrix $u \cdot v \in \mathcal{F}_k$ as follows. For $1 \leq i, j \leq k$,

$$(u \cdot v)_{i,j} = \begin{cases} u_i v_j & \text{if } i \neq j, \\ 0 & \text{otherwise}. \end{cases}$$

Thus $u \cdot v$ differs from the Kronecker product $u^T \otimes v$ just by having zeroes on the main diagonal. Clearly, $u \cdot v = (v \cdot u)^T$.

3. The fairness of DOL systems

We will prove here that the fairness of a DOL sequence over a $k$-letter alphabet is determined by the fairness of its $k + 1$ initial elements (Theorem 8). Our proof is based on the observation of the ways a two-symbol subword may occur in a morphic image (in the general morphism $h : \Sigma^* \rightarrow A^*$) of a word.

**Lemma 2.** Let $w \in \Sigma^*$. Then, for $0 \leq r, s \leq m$, $0 \leq i, j \leq k$,

$$|h(w)|_{b_i, b_j} = \sum_i |w|_{ai} |h(ai)|_{b_i} + \sum_{i,j} |w|_{ai,aj} |h(ai)|_{b_i} |h(aj)|_{b_j}.$$

**Proof.** There are two ways a subword $b_i b_j$ may be contained in $h(w)$: (1) As a subword of a single factor $h(ai)$—this happens $|h(ai)|_{b_i}$ times in such a factor and there are $|w|_{ai}$ factors $h(ai)$ for the particular symbol $a_i$. (2) The symbol $b_i$ occurs in a factor $h(ai)$ and the symbol $b_j$ in a consecutive factor $h(aj)$. There are $|h(ai)|_{b_i}$ occurrences of $b_i$ in the first factor and $|h(aj)|_{b_j}$ occurrences of $b_j$ in the second factor and there are $|w|_{ai,aj}$ such pairs of factors for the particular symbols $a_i$ and $a_j$.

Denote $p^\Sigma = P_w^\Sigma$, $p^D = P_w^D$, and, for $1 \leq r \leq k$, $p^{(r)} = p_{h(ai)}^D$, $\Pi^{(r)} = P_{h(ai)}^D$. The following theorem re-states the assertion of Lemma 2 in the terms of Parikh vectors and $p$-matrices.
Theorem 3.

$$P_{h(w)}^A = \sum_i p_i \Pi^{(i)} + \sum_{i,j \atop i \neq j} P_{i,j}(\pi^{(i)} \cdot \pi^{(j)}) + \sum_i \left( \frac{p_i}{2} \right)(\pi^{(i)} \cdot \pi^{(i)}),$$

where $1 \leq i, j \leq k$.

**Remark 4.** Since $\pi^{(i)} \cdot \pi^{(j)} = \Pi^{(i)} \circ \Pi^{(j)} - (\Pi^{(i)} + \Pi^{(j)})$, $P_{h(w)}^A$ can be expressed as

$$P_{h(w)}^A = \sum_i P_{i,i} \Pi^{(i)} + \sum_{i,j \atop i \neq j} P_{i,j}(\Pi^{(i)} \circ \Pi^{(j)} - (\Pi^{(i)} + \Pi^{(j)})) + \sum_i \left( \frac{p_i}{2} \right)(\Pi^{(i)} \circ \Pi^{(i)} - 2\Pi^{(i)}).$$

In [9] the difference function was introduced for a binary alphabet. We present here balance matrices as a natural generalization of this function for alphabets of any size. The balance matrices are closely related to $p$-matrices. Let $P_w$ be the $p$-matrix of a word $w$. The balance matrix (b-matrix, for short) of $w$ is the matrix $B_w^\Sigma = P_w^\Sigma - (P_w^\Sigma)^T$. Hence $B_w^\Sigma = P_w^\Sigma - P_{w,R}$. $(B_w^\Sigma)^T = -B_w^\Sigma$ and $B_w^\Sigma = \mathbf{0}^k$ iff $w$ is a fair word. Denote $B = B_w^\Sigma$, and, for $1 \leq r \leq k$, $\Theta^{(r)} = B_{h(\alpha_r)}^A$. The following Theorem 5, Corollary 6 and Theorem 8 are generalizations of Theorem 6, Lemma 1 and Theorem 2 from [8], respectively.

**Theorem 5.**

$$B_{h(w)}^A = \sum_i p_i \Theta^{(i)} + \sum_{i \neq j} B_{i,j}(\pi^{(i)} \cdot \pi^{(j)} - \pi^{(j)} \cdot \pi^{(i)}).$$

**Proof.** The assertion follows Theorem 3, since $(\pi^{(i)} \cdot \pi^{(j)})^T = \pi^{(j)} \cdot \pi^{(i)}$. \hfill \Box

**Corollary 6.** If $w$ is a fair word then $B_{h(w)}^A = \sum_i p_i \Theta^{(i)}$.

**Remark 7.** Using Remark 4, we obtain

$$B_{h(w)}^A = \sum_i p_i \Theta^{(i)} + \sum_{i \neq j} B_{i,j}(\Pi^{(i)} \circ \Pi^{(j)} - \Pi^{(j)} \circ \Pi^{(i)}).$$

**Theorem 8.** A DOL sequence over a k-letter alphabet is fair if and only if the initial $k + 1$ words in the sequence are fair.

**Proof.** Let $G = (\Sigma, h, w)$ be a DOL system. We may assume that $w$ is not the empty word. Denote, for $s = 0, 1, \ldots$, $p^{(s)} = p_{h(w)}^A$, $B^{(s)} = B_{h(w)}^A$. First observe, that, for $s \geq 0$, $p^{(s+1)} = p^{(s)} \Psi$, hence $p^{(s)} = p^{(0)} \Psi^s$, where $\Psi = ((\pi^{(1)})^T, \ldots, (\pi^{(k)})^T)^T$. We will prove that $B^{(s)} = \mathbf{0}^k$ by induction on $s$. By the assumption of the theorem, the assertion is true for $0 \leq s \leq k$. Let us now assume that, for some $s \geq k + 1$, $B^{(s)} = \mathbf{0}^k$ for all values $0 \leq t \leq s - 1$. The $k + 1$-dimensional vectors $p^{(0)}, \ldots, p^{(k)}$ are linearly dependent, hence, for some $1 \leq r \leq k$ there exist...
rational numbers $q_0, \ldots, q_{r-1}$ such that $p^{(r)} = \sum_{j=0}^{r-1} q_j p^{(j)}$. Then

$$p^{(s-1)} = p^{(0)} \Psi^{s-1} = p^{(0)} \Psi^{s-1-r} = \left( \sum_{j=0}^{r-1} q_j p^{(0)} \Psi^j \right) \Psi^{s-1-r} = \sum_{j=0}^{r-1} q_j p^{(0)} \Psi^{s-1-r+j} = \sum_{j=0}^{r-1} q_j p^{(s-1-r+j)}.$$

Corollary 6 implies that

$$B^{(s)} = \sum_i p_i^{(s-1)} \Theta^{(i)} = \sum_i \left( \sum_{j=0}^{r-1} q_j p_i^{(s-1-r+j)} \right) \Theta^{(i)} = \sum_i \sum_{j=0}^{r-1} q_j p_i^{(s-1-r+j)} \Theta^{(i)} = \sum_{j=0}^{r-1} q_j \sum_i p_i^{(s-1-r+j)} \Theta^{(i)} = \sum_{j=0}^{r-1} q_j B^{(s-r+j)} = 0^{(k)}.$$

**Corollary 9.** If $k + 1$ consecutive words in a D0L sequence over a $k$-letter alphabet are fair, then all words following them in the sequence are fair.

**Example 10.** Consider the D0L system $([a, b, c], h, a)$ where $h(a) = acb$, $h(b) = bca$, $h(c) = c$. Then $S(G)$ is the sequence with $c$ at every odd position (positions are numbered starting from 0) and the blocks of the Thue sequence on the alphabet $\{a, b\}$ in the even positions:

$$a, acb, acbcaacacbcaacb, acbcaacacbcaacacbaacbcacbcaacb, \ldots.$$

The $p$-matrices corresponding to the first 6 words in the sequence are

$$
\begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 
\end{pmatrix},
\begin{pmatrix}
1 & 1 & 1 \\
0 & 1 & 0 \\
0 & 1 & 1 
\end{pmatrix},
\begin{pmatrix}
2 & 2 & 3 \\
2 & 2 & 3 \\
3 & 3 & 3 
\end{pmatrix},
\begin{pmatrix}
4 & 8 & 14 \\
8 & 4 & 14 \\
14 & 14 & 7 
\end{pmatrix},
\begin{pmatrix}
8 & 32 & 60 \\
32 & 8 & 60 \\
60 & 60 & 15 
\end{pmatrix},
\begin{pmatrix}
16 & 128 & 248 \\
128 & 16 & 248 \\
248 & 248 & 31 
\end{pmatrix}.
$$

The words at positions 2, 3, 4 and 5 are fair, hence all words following in the sequence are fair.
Acknowledgment

The author is thankful to A. Salomaa for several valuable comments and suggestions.

References