Discrete Applied Mathematics 161 (2013) 2851–2861

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Counting minimal semi-Sturmian words[☆]

F. Blanchet-Sadri[a,*], Sean Simmons[b]

[a] Department of Computer Science, University of North Carolina, P.O. Box 26170, Greensboro, NC 27402–6170, USA

[b] Department of Mathematics, Massachusetts Institute of Technology, Building 2, Room 236, 77 Massachusetts Avenue, Cambridge, MA 02139–4307, USA

ARTICLE INFO

Article history:
Received 24 December 2011
Received in revised form 3 July 2013
Accepted 5 July 2013
Available online 25 July 2013

Keywords:
Combinatorics on words
Graph theory
Subword complexity
Semi-Sturmian words
Euler's totient function

ABSTRACT

A finite *Sturmian* word w is a balanced word over the binary alphabet $\{a, b\}$, that is, for all subwords u and v of w of equal length, $||u|_a - |v|_a| \leq 1$, where $|u|_a$ and $|v|_a$ denote the number of occurrences of the letter a in u and v, respectively. There are several other characterizations, some leading to efficient algorithms for testing whether a finite word is Sturmian. These algorithms find important applications in areas such as pattern recognition, image processing, and computer graphics. Recently, Blanchet-Sadri and Lensmire considered finite *semi-Sturmian* words of minimal length and provided an algorithm for generating all of them using techniques from graph theory. In this paper, we exploit their approach in order to count the number of minimal semi-Sturmian words. We also present some other results that come from applying this graph theoretical framework to subword complexity.

1. Introduction

An infinite word w is an infinite sequence of letters from a finite alphabet. Any finite block of consecutive letters of w is a *factor* or *subword* of w. The word w is *Sturmian* if, for all non-negative integers n, there are exactly $n+1$ distinct subwords of w of length n. In other words, the *subword complexity* $p_w(n)$ of w, which counts the number of distinct subwords of length n of w, is equal to $n+1$. The fact $p_w(1) = 1 + 1 = 2$ implies that w is constructed from two distinct letters of the alphabet. Without loss of generality, we call these a and b. The well-known Fibonacci word

$$abaababaabaababaababaabaababaabaab\ldots$$

is Sturmian. It is defined by $F_{n+2} = F_{n+1}F_n$, where $F_0 = a$ and $F_1 = ab$.

Sturmian words have been widely studied. Morse and Hedlund introduced the term "Sturmian trajectories" and did a first comprehensive study in 1940 in relation to symbolic dynamics [13]. Chapter 2 of Lothaire's book "Algebraic Combinatorics on Words" provides a systematic exposition of Sturmian words, their numerous properties, and equivalent definitions [11]. Sturmian words appear in the literature under various names: rotation sequences, cutting sequences, Christoffel words, Beatty sequences, characteristic words, balanced words, nonhomogeneous spectra, billiard trajectories, etc. Application areas include linear filters [10], routing in networks [1], pattern recognition [5], image processing and computer graphics [6]. For example, counting the number of distinct digitized straight lines corresponds to counting the number of subwords of a given length in Sturmian words. A formula was conjectured by Dulucq and Gouyou-Beauchamps in [8] and later proved by Mignosi in [12].

☆ This material is based upon work supported by the National Science Foundation under Grant No. DMS–1060775. We thank the referees of a preliminary version of this paper for their very valuable comments and suggestions.

* Corresponding author. Tel.: +1 3368550722.
E-mail addresses: blanchet@uncg.edu (F. Blanchet-Sadri), seanken@mit.edu (S. Simmons).

0166-218X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dam.2013.07.008

A finite word w is Sturmian if it is a subword of an infinite Sturmian word. Linear-time algorithms have been provided for recognizing finite Sturmian words (see for example, Boshernitzan and Fraenkel [4] and de Luca and De Luca [7]). Berstel and Pocchiola also provided a linear probabilistic algorithm for generating randomly finite Sturmian words [2].

Now, a finite word w is *semi-Sturmian of order N* if $p_w(n) = n+1$ for $n = 1, \ldots, N$. Note that the terminology *Sturmian of order N* was previously used by Blanchet-Sadri and Lensmire in [3] for such word, but we decided to adopt the terminology "semi-Sturmian" here to avoid confusion with finite Sturmian words. Not all semi-Sturmian words of order N are Sturmian, for instance, *aabb* is a semi-Sturmian word of order 2 but it is not a subword of any infinite Sturmian word. However every finite Sturmian word is semi-Sturmian of order N for some N. A semi-Sturmian word of order N is *minimal* if it has minimal length among all semi-Sturmian words of order N. Equivalently, it is minimal if it has length $2N$. In [3], Blanchet-Sadri and Lensmire described an algorithm that generates all minimal semi-Sturmian words of each order $N \geq 3$. Earlier in [14], it had been shown that the minimal length of a word w such that $p_w(n) = F_{n+2}$ for all n, $1 \leq n \leq N$, is $F_N + F_{N+2}$, where $(F_n)_{n\geq 1}$ is the Fibonacci sequence and N is a positive integer, and an algorithm had been given for generating such minimal words of each order $N \geq 1$.

In this paper, our main result is to count the number of minimal semi-Sturmian words of order N for every integer N greater than 1. We show that this number is connected to Euler's totient function ϕ from number theory, where the totient $\phi(n)$ of a positive integer n is the number of positive integers less than or equal to n that are coprime to n.

The contents of our paper is as follows. In Section 2, we review some basics on semi-Sturmian graphs and some graphs corresponding to given sets of words of a fixed length. We also recall conditions for the existence of Eulerian paths in graphs. In Section 3, we consider minimal words with subword complexity $n+1$, that is, minimal semi-Sturmian words. We count all minimal semi-Sturmian words of order N using a graph theoretical approach based on the above mentioned algorithm that generates all such words. We show that any graph produced by this algorithm belongs to one of three families of semi-Sturmian graphs that end up playing an important role in the counting. In Section 4, we use our techniques to extend our result further to include a lower bound on the number of minimal words with subword complexity $n + k - 1$, where k is the alphabet size.

2. Preliminaries on graphs

We recall some graph theoretical concepts that will be useful. All graphs in this paper are assumed to be directed. The reader is referred to [9] for more information.

A graph G is said to be *semi-Sturmian of order n* if G has n vertices, $n+1$ edges, and contains an Eulerian path. The graph G is also said to be *semi-Sturmian* if it is semi-Sturmian of some order n. Moreover for any graph $G = (V, E)$, we denote by $L(G)$ its *line graph* which is the graph $G' = (V', E')$ where $V' = E$, and for all $v_1', v_2' \in V'$, $(v_1', v_2') \in E'$ if $v_1' = (v_1, v_2)$ and $v_2' = (v_2, v_3)$ for some $v_1, v_2, v_3 \in V$. Fig. 1(c) gives the line graph of Fig. 1(b).

Now, let S be a set of words of length n. Combining ideas from de Bruijn and Rauzy graphs, the graph $G_S = (V, E)$, defined in [3], is as follows: V is the set of all factors of length $n-1$ of words in S, and E consists of all edges (x, x') so that there exists a word $y \in S$ with x as a prefix and x' as a suffix. The edge (x, x') can be identified (or labelled) with the word y. See Fig. 1(a) for an example where $S = \{aa, ab, ba, bb\}$.

It is worth noting that every path in a graph of the form G_S corresponds to a word. More specifically, let $x_0, \ldots, x_m$ be a path in G_S where $x_0, \ldots, x_m$ are vertices. Then this path corresponds to the word w where $w[0 \ldots n-1) = x_0$, $w[1 \ldots n) = x_1, \ldots, w[|w| - n + 1 \ldots |w|) = x_m$ (here $m = |w| - n + 1$). Moreover if p and q are different paths, then they correspond to different words. A similar construction allows us to view every path in the graphs $L(G_S), L(L(G_S)), \ldots$ as a word. We say that if p is a path in some subgraph of $L(\cdots(L(G_S))\cdots)$, then p corresponds to a word.

We end this section with a well-known result on the existence of Eulerian paths. The notation $\texttt{ideg}(v)$ refers to the indegree of vertex v and $\texttt{odeg}(v)$ to its outdegree.

Lemma 1. *Let G be a graph, and let x and y be vertices in G.*

- *If $x = y$, then there is an Eulerian path from x to y if and only if G is strongly connected and* $\texttt{ideg}(v) = \texttt{odeg}(v)$ *for all v.*
- *If $x \neq y$, then there is an Eulerian path from x to y if and only if G is weakly connected,* $\texttt{ideg}(x) = \texttt{odeg}(x) - 1$, $\texttt{ideg}(y) = \texttt{odeg}(y) + 1$, *and* $\texttt{ideg}(v) = \texttt{odeg}(v)$ *for all other vertices v.*

3. Our main result

Our main goal is to prove the following result. Recall that the *Euler totient* $\phi(n)$ of a positive integer n is the number of positive integers less than or equal to n that are coprime to n. For example, $\phi(9) = 6$ since 1, 2, 4, 5, 7 and 8 are coprime to 9.

Theorem 1. *For $N \geq 2$, the number of minimal semi-Sturmian words of order N, $S(N)$, satisfies*

$$S(N) = 6 + 4\left(\sum_{n=3}^{N} \phi(n)\right) + (N-1)\phi(N+1)$$

where $\phi(n)$ is the Euler totient function.

We begin by recalling an algorithm due to Blanchet-Sadri and Lensmire [3], which we illustrate in Fig. 1.