Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

On the density of Lyndon roots in factors

Maxime Crochemore [a,b], Robert Mercaş [a,c,*]

[a] King's College London, United Kingdom
[b] Université Paris-Est, France
[c] Kiel University, Germany

ARTICLE INFO

Article history:
Received 19 August 2015
Received in revised form 3 February 2016
Accepted 12 February 2016
Available online xxxx

Keywords:
Runs
Lyndon words
Lyndon roots
Critical factorisation

ABSTRACT

This work takes another look at the number of runs that a string may contain and provides an alternative proof for the bound. We also propose another stronger conjecture that states the following: for a fixed order on the alphabet, within every factor of a word there are at most as many occurrences of Lyndon roots corresponding to runs in the word as the length of the factor. Only first occurrences of roots in each run are considered.

1. Introduction

The concept of a run coined by Iliopoulos et al. [11] when analysing repetitions in Fibonacci words, has been introduced to represent in a succinct manner all occurrences of repetitions in a word. It is known that there are only $\mathcal{O}(n)$ many of them in a word of length n from Kolpakov and Kucherov [12] who proved it in a non-constructive manner. The first explicit bound was later on provided by Rytter [15]. Several improvements on the upper bound can be found in [16,4,14,5,8]. Kolpakov and Kucherov conjectured that this number is in fact smaller than n, which has been proved by Bannai et al. [1,2]. Recently, Holub [10] and Fischer et al. [9] gave a tighter upper bound reaching $22n/23$.

In this note we provide a proof of the result, slightly different than the short and elegant proof in [2]. Then we provide a relation between the border-free root conjugates of a square and the critical positions [13, Chapter 8] occurring in it. Finally, counting runs extends naturally to the question of their highest density, that is, to the question of the type of factors in which there is a large accumulation of runs. This is treated in the last section.

Formally, a *run* in a word w is an interval $[i\,..\,j]$ of positions, $0 \le i < j < |w|$, for which both the associated factor $w[i\,..\,j]$ is periodic (i.e. its smallest period p satisfies $p \le (j-i+1)/2$), and the periodicity cannot be extended to the right nor to the left: $w[i-1\,..\,j]$ and $w[i\,..\,j+1]$ have larger periods when these words are defined (see Fig. 1).

2. Fewer runs than length

We consider an ordering $<$ on the word alphabet and the corresponding lexicographic ordering denoted $<$ as well. We also consider the lexicographic ordering $\tilde{<}$, called the reverse ordering, inferred by the inverse alphabet ordering $<^{-1}$. Recall that, for a fixed ordering on the alphabet, a Lyndon word is a primitive word that is not greater than any of its conjugates

* Corresponding author.
E-mail addresses: Maxime.Crochemore@kcl.ac.uk (M. Crochemore), robertmercas@gmail.com (R. Mercaş).

http://dx.doi.org/10.1016/j.tcs.2016.02.015
0304-3975/

Please cite this article in press as: M. Crochemore, R. Mercaş, On the density of Lyndon roots in factors, Theoret. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.tcs.2016.02.015

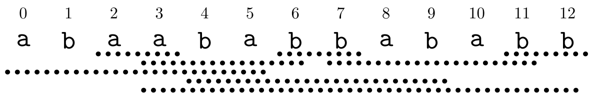

Fig. 1. Dotted lines show the 8 runs in abaababbababb. For example, $[7\,..\,11]$ is the run of period 2 and length 5 associated with factor babab.

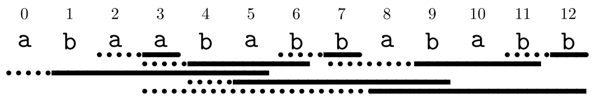

Fig. 2. Plain lines show the 8 greatest proper suffixes assigned to runs of abaababbababb from Fig. 1 in the proof of the theorem. Note that no two suffixes start on the same position.

(rotations). Equivalently, it is smaller than all its proper suffixes. The main element in the proof is to assign to each run its (lexicographically) greatest suffix according to one of the two orderings (see Fig. 2).

Theorem 1. *The number of runs in a word of length n is less than n.*

Proof. Let w be a word of length n. Let $[i\,..\,j]$ $(0 \le i < j < n)$ be a run of smallest period p in w. If $j+1 < n$ and $w[j+1] > w[j-p+1]$ we assign to the run the position k for which $w[k\,..\,j]$ is the greatest proper suffix of $w[i\,..\,j]$. Else, k is the position of the greatest proper suffix of $w[i\,..\,j]$ according to $\tilde{<}$.

Note that if $k > i$ then $k > 0$, and that $w[k\,..\,j]$ contains a full period of the run factor, i.e. $j-k+1 \ge p$. Also note that $w[k\,..\,k+p-1]$ is a greatest conjugate of the period root $w[i\,..\,i+p-1]$ according to one of the two orderings. Therefore, it is border-free, which is a known property of Lyndon words.

We claim that each position $k > 0$ on w is the starting position of at most one greatest proper suffix of a run factor. Let us consider two distinct runs $[i\,..\,j]$ and $[\bar{i}\,..\,\bar{j}]$ of respective periods p and q, and which are called respectively the p-run and the q-run. For the sake of contradiction, we assume that their greatest suffixes share the same starting position k. Assume $p \ne q$ since the runs cannot be distinct and have the same period.

First case, $j = \bar{j}$, which implies $w[k\,..\,j] = w[k\,..\,\bar{j}]$. Assume for example that $p < q$. Then, $w[k\,..\,k+q-1]$ has period p and thus is not border-free, which is a contradiction.

Second case, assume without loss of generality that $j < \bar{j}$ and that both suffixes are the greatest in their runs according to the same ordering, say $<$. Let $d = w[j+1]$, the letter following the p-run. By definition we have $w[j-p+1] < d$ and then $w[i\,..\,j-p+1] < w[i\,..\,j-p]d$. But since $w[i+p\,..\,j]d$ is a factor of the q-run this contradicts the maximality of $w[k\,..\,\bar{j}-1]$.

Third case, $j \ne \bar{j}$ and the suffixes are greatest according to different orderings. Assume without loss of generality that $p < q$ and the suffix of the p-run factor is greatest according to $<$. Since $q > 1$ we have both $w[k+q-1] \tilde{>} w[k]$ and $w[k+q-1] = w[k-1]$, then $w[k-1] < w[k]$. We cannot have $p > 1$ because this implies $w[k-1] > w[k]$. And we cannot have either $p = 1$ because this implies $w[k-1] = w[k]$. Therefore we get again a contradiction.

This ends the proof of the claim and shows that the number of runs is no more than the number $n-1$ of potential values for k, as stated. □

3. Lyndon roots

The proof of Theorem 1 by Bannai et al. [2] relies on the notion of a Lyndon root. The root of a run $[i\,..\,j]$ of period p in w is the factor $w[i\,..\,i+p-1]$. Henceforth, the Lyndon root of a run is the Lyndon conjugate of its root. Therefore, since a run has length at least twice as long as its root, the first occurrence of its Lyndon root is followed by its first letter. This notion of Lyndon root is the basis of the proof of the $0.5n$ upper bound on the number of cubic runs given in [6]. Recall that a run is said to be cubic if its length is at least three times larger than its period.

Lyndon roots considered in [2] are defined according to the two orderings $<$ and $\tilde{<}$. However, these Lyndon roots can be defined as smallest or greatest conjugates of the run root according to only one ordering.

The proof of Theorem 1 is inspired by the proof in [2] but does not use explicitly the notion of Lyndon roots. The link between the two proofs is as follows: when the suffix $w[k\,..\,j]$ is greatest according to $<$ in the run factor, then its prefix of period length, $w[k\,..\,k+p-1]$, is a Lyndon word according to $\tilde{<}$. As a consequence, the assignment of positions to runs is almost the same whatever greatest suffixes or Lyndon roots are considered.

The use of Lyndon roots leaves more flexibility to assign positions to runs. Indeed, a run factor may contain several occurrences of the run's Lyndon root. Furthermore, any two consecutive occurrences of this root do not overlap and are adjacent. The multiplicity of these occurrences can be transposed to greatest suffixes by considering their borders. Doing so, what is essential in the proof of Theorem 1 is that the suffixes and borders so defined are at least as long as the period of the run. Consequently, consecutive such marked positions can be assigned to the same run. As a consequence, since every cubic run is associated to at least two positions, this yields the following corollaries.

Please cite this article in press as: M. Crochemore, R. Mercaş, On the density of Lyndon roots in factors, Theoret. Comput. Sci. (2016), http://dx.doi.org/10.1016/j.tcs.2016.02.015