Advances in Applied Mathematics 101 (2018) 281–319

Contents lists available at ScienceDirect

Advances in Applied Mathematics

www.elsevier.com/locate/yaama

Inverse Lyndon words and inverse Lyndon factorizations of words

Paola Bonizzoni [a], Clelia De Felice [b,*], Rocco Zaccagnino [b], Rosalba Zizza [b]

[a] *Dipartimento di Informatica Sistemistica e Comunicazione, Università degli Studi di Milano Bicocca, Viale Sarca 336, 20126 Milano, Italy*
[b] *Dipartimento di Informatica, Università degli Studi di Salerno, via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy*

ARTICLE INFO

Article history:
Received 31 July 2018
Received in revised form 2 August 2018
Accepted 12 August 2018
Available online xxxx

MSC:
68R15
68W32

Keywords:
Lyndon words
Lyndon factorization
Combinatorial algorithms on words
DNA sequences

ABSTRACT

Motivated by applications to string processing, we introduce variants of the Lyndon factorization called inverse Lyndon factorizations. Their factors, named inverse Lyndon words, are in a class that strictly contains anti-Lyndon words, that is Lyndon words with respect to the inverse lexicographic order. The Lyndon factorization of a nonempty word w is unique but w may have several inverse Lyndon factorizations. We prove that any nonempty word w admits a canonical inverse Lyndon factorization, named ICFL(w), that maintains the main properties of the Lyndon factorization of w: it can be computed in linear time, it is uniquely determined, and it preserves a compatibility property for sorting suffixes. In particular, the compatibility property of ICFL(w) is a consequence of another result: any factor in ICFL(w) is a concatenation of consecutive factors of the Lyndon factorization of w with respect to the inverse lexicographic order.

* Corresponding author.
E-mail addresses: bonizzoni@disco.unimib.it (P. Bonizzoni), cdefelice@unisa.it (C. De Felice), zaccagnino@dia.unisa.it (R. Zaccagnino), rzizza@unisa.it (R. Zizza).

https://doi.org/10.1016/j.aam.2018.08.005
0196-8858/

1. Introduction

Lyndon words were introduced in [34], as *standard lexicographic sequences*, and then used in the context of the free groups in [8]. A Lyndon word is a word which is strictly smaller than each of its proper cyclic shifts for the lexicographical ordering. A famous theorem concerning Lyndon words asserts that any nonempty word factorizes uniquely into a nonincreasing product of Lyndon words, called its Lyndon factorization. This theorem, that can be recovered from results in [8], provides an example of a factorization of a free monoid, as defined in [42] (see also [4,32]). Moreover, there are several results which give relations between Lyndon words, codes and combinatorics of words [3]. More recently these words found a renewed theoretical interest and several variants of them have been studied [7,16]. A related field studies the combinatorial and algorithmic properties of *necklaces*, that are powers of Lyndon words, and their prefixes or *prenecklaces* [6].

The Lyndon factorization has recently revealed to be a useful tool also in string processing algorithms [2,38] with a potential that has not been completely explored and understood. This is due also to the fact that it can be efficiently computed. Linear-time algorithms for computing this factorization can be found in [17,18] whereas an $\mathcal{O}(\lg n)$-time parallel algorithm has been proposed in [1,14]. A connection between the Lyndon factorization and the Lempel–Ziv (LZ) factorization has been given in [26], where it is shown that in general the size of the LZ factorization is larger than the size of the Lyndon factorization, and in any case the size of the Lyndon factorization cannot be larger than a factor of 2 with respect to the size of LZ.

A Lyndon word is lexicographically smaller than all its proper nonempty suffixes. This explains why the Lyndon factorization has become of particular interest also in suffix sorting problems. The suffix array (SA) of a word w is the lexicographically ordered list of the starting positions of the suffixes of w. The connection between Lyndon factorizations and suffix arrays has been pointed out in [25], where the authors show a method to construct the Lyndon factorization of a text from its SA. Conversely, the computation of the SA of a text from its Lyndon factorization has been proposed in [5] and then explored in [36,37].

The algorithm proposed in [36,37] is based on the following interesting combinatorial result, proved in the same papers: if u is a concatenation of consecutive Lyndon factors of $w = xuy$, then the position of a nonempty suffix u_i in the ordered list of suffixes of u (called *local* suffixes) is the same as the position of the nonempty suffix $u_i y$ in the ordered list of the suffixes of w (called *global* suffixes). In turn, this result suggests a divide and conquer strategy for the sorting of the suffixes of a word $w = w_1 w_2$: we order the nonempty suffixes of w_1 and the nonempty suffixes of w_2 independently (or in parallel) and then we merge the resulting lists (see Section 2.4 for further details).

Relations between Lyndon words and the Burrows–Wheeler Transform (BWT) have been discovered first in [11,35] and, more recently, in [29] and in [23]. Variants of the BWT