International Workshop On
Combinatorial Algorithms
IWOCA
http://www.iwoca.org → Problems Section

Maximum Number of Distinct Lyndon Subsequences

Dominik Köppl

Tokyo Medical and Dental University, Japan,
koeppl.dsc@tmd.ac.jp

presented at IWOCA 2022 on June 10
last update: November 2, 2022

Abstract

For a fixed length n and an alphabet Σ, what is the maximum number of distinct Lyndon subsequences a string of length n can have?

Definition 1. A string is called *Lyndon* if it is strictly lexicographically smaller than all its proper suffixes. Alternatively, it is Lyndon if it is strictly lexicographically smaller than all its cyclic rotations. For instance, a, ab, $aabab$ are Lyndon, but neither aa, $abab$, nor $abaab$.

Definition 2. A subsequence of a string $T[1..n]$ of the length ℓ is a string of the form $T[i_1] \cdot T[i_2] \cdots T[i_\ell]$ with $1 \leq i_1 < i_2 < \cdots < i_\ell \leq n$.

Problem 1 ([2])**.** What is the maximum number of distinct Lyndon subsequences in a string of length n over an alphabet of size σ?

Comment 1. Trivial cases are $\sigma \in \{1, n\}$. For $\sigma = 1$, the string is unary $T = a \dots a$, and therefore has only one distinct Lyndon sequence, namely a. For $\sigma = n$, we can enumerate the characters by their ranks from 1 to n, and study $T = 1 \cdot 2 \cdot 3 \cdots n$, for which we can see that any subsequence forms a Lyndon subsequence. Since the number of subsequences is 2^n for a string of length n, the answer to our problem is also 2^n.

Some results we achieved at the 4th AFSA SSSS 2022 [3]:

Comment 2. We can interpret the string $T = 1 \cdot 2 \cdot 3 \cdots m$ with $m = 2^d$ as a bit vector B of length $n := md = m \lg m$. Since T has 2^m distinct Lyndon subsequences, so has B by taking always blocks of length d. So we have at least $2^m = 2^{n/\lg m} = \Theta(2^{n/\lg n})$ different Lyndon subsequences in B.

Comment 3. For a given k, consider the string $T = P \cdot \prod_{j=1}^{x-1}(B_j \cdot 1)$ on the binary alphabet $\{0, 1\}$, where $P = 0 \cdots 0$ is a run of zeros of length k, and B_j an arbitrary string of length $k-1$ for $j \in [1..x-1]$. Then any subsequence of

T of the form $P \cdot \prod_{j=1}^{x-1}(B'_j \cdot 1)$ is Lyndon, where B'_j is a subsequence of B_j. If we say that the length of T is n, then the number of characters of all $B'_j s$ is $\frac{x-1}{x}n - (x-1)$. This number is maximized to $n - 2\sqrt{n} + 1$ when $x = \sqrt{n}$. Consequently, we can select $n - 2\sqrt{n} + 1$ characters at random. According to [1], the maximum number of distinct subsequences is given by $(n+3)$-th Fibonacci number decremented by one, for a string of length n. In our case, this gives a new lower bound of about $(1.618)^{n-2\sqrt{n}-3}/\sqrt{5}$.

Problem 2. Still countable are all Lyndon subsequences of T of length at most $k = \sqrt{n}$ since we have only counted those that start with P. There are at least $\sqrt{n}$ many of the form $0 \cdots 01 \cdots 1$, but probably more.

References

[1] Abraham Flaxman, Aram W. Harrow, and Gregory B. Sorkin. Strings with maximally many distinct subsequences and substrings. *Electron. J. Comb.*, 11(1), 2004. `doi:10.37236/1761`.

[2] Ryo Hirakawa, Yuto Nakashima, Shunsuke Inenaga, and Masayuki Takeda. Counting Lyndon subsequences. In *Proc. PSC*, pages 53–60, 2021. URL: `http://www.stringology.org/event/2021/p05.html`.

[3] Takashi Horiyama, Dominik Koeppl, Shinichi Minato, Hirotaka Ono, Toshiki Saitoh, Ryuhei Uehara, and Yushi Uno. 4th meeting of AFSA group B01, 2022.