Counting Lyndon Subsequences

Ryo Hirakawa[1], Yuto Nakashima[2], Shunsuke Inenaga[2,3], and Masayuki Takeda[2]

[1] Department of Information Science and Technology, Kyushu University, Fukuoka, Japan
hirakawa.ryo.460@s.kyushu-u.ac.jp
[2] Department of Informatics, Kyushu University, Fukuoka, Japan
{yuto.nakashima, inenaga, takeda}@inf.kyushu-u.ac.jp
[3] PRESTO, Japan Science and Technology Agency, Kawaguchi, Japan

Abstract. Counting substrings/subsequences that preserve some property (e.g., palindromes, squares) is an important mathematical interest in stringology. Recently, Glen et al. studied the number of Lyndon factors in a string. A string $w = uv$ is called a Lyndon word if it is the lexicographically smallest among all of its conjugates vu. In this paper, we consider a more general problem "counting Lyndon subsequences". We show (1) the maximum total number of Lyndon subsequences in a string, (2) the expected total number of Lyndon subsequences in a string, (3) the expected number of distinct Lyndon subsequences in a string.

1 Introduction

A string $x = uv$ is said to be a *conjugate* of another string y if $y = vu$. A string w is called a *Lyndon word* if it is the lexicographically smallest among all of its conjugates. It is also known that w is a Lyndon word iff w is the lexicographically smallest suffix of itself (excluding the empty suffix).

A *factor* of a string w is a sequence of characters that appear contiguously in w. A factor f of a string w is called a *Lyndon factor* if f is a Lyndon word. Lyndon factors enjoy a rich class of algorithmic and stringology applications including: counting and finding the maximal repetitions (a.k.a. runs) in a string [2] and in a trie [8], constant-space pattern matching [3], comparison of the sizes of run-length Burrows-Wheeler Transform of a string and its reverse [4], substring minimal suffix queries [1], the shortest common superstring problem [7], and grammar-compressed self-index (Lyndon-SLP) [9].

Since Lyndon factors are important combinatorial objects, it is natural to wonder how many Lyndon factors can exist in a string. Regarding this question, the next four types of counting problems are interesting:

- $MTF(\sigma, n)$: the *maximum total* number of Lyndon factors in a string of length n over an alphabet of size σ.
- $MDF(\sigma, n)$: the *maximum* number of *distinct* Lyndon factors in a string of length n over an alphabet of size σ.
- $ETF(\sigma, n)$: the *expected total* number of Lyndon factors in a string of length n over an alphabet of size σ.
- $EDF(\sigma, n)$: the *expected* number of *distinct* Lyndon factors in a string of length n over an alphabet of size σ.

Glen et al. [5] were the first who tackled these problems, and they gave exact values for $MDF(\sigma, n)$, $ETF(\sigma, n)$, and $EDF(\sigma, n)$. Using the number $L(\sigma, n)$ of Lyndon words of length n over an alphabet of size σ, their results can be written as shown in Table 1.

Ryo Hirakawa, Yuto Nakashima, Shunsuke Inenaga, Masayuki Takeda: Counting Lyndon Subsequences, pp. 53–60.
Proceedings of PSC 2021, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-06869-4

Number of Lyndon Factors in a String	
Maximum Total $MTF(\sigma, n)$	$\binom{n+1}{2} - (\sigma - p)\binom{m+1}{2} - p\binom{m+2}{2} + n$ [this work]
Maximum Distinct $MDF(\sigma, n)$	$\binom{n+1}{2} - (\sigma - p)\binom{m+1}{2} - p\binom{m+2}{2} + \sigma$ [5]
Expected Total $ETF(\sigma, n)$	$\sum_{m=1}^{n} L(\sigma, m)(n - m + 1)\sigma^{-m}$ [5]
Expected Distinct $EDF(\sigma, n)$	$\sum_{m=1}^{n} L(\sigma, m) \sum_{s=1}^{\lfloor n/m \rfloor} (-1)^{s+1} \binom{n - sm + s}{s} \sigma^{-sm}$ [5]

Table 1. The numbers of Lyndon factors in a string of length n over an alphabet of size σ, where $n = m\sigma + p$ with $0 \le p < \sigma$ for $MTF(\sigma, n)$ and $MDF(\sigma, n)$.

The first contribution of this paper is filling the missing piece of Table 1, the exact value of $MTF(\sigma, n)$, thus closing this line of research for Lyndon factors (substrings).

We then extend the problems to subsequences. A subsequence of a string w is a sequence of characters that can be obtained by removing 0 or more characters from w. A subsequence s of a string w is said to be a *Lyndon subsequence* if s is a Lyndon word. As a counterpart of the case of Lyndon factors, it is interesting to consider the next four types of counting problems of Lyndon subsequences:

- $MTS(\sigma, n)$: the *maximum total* number of Lyndon subsequences in a string of length n over an alphabet of size σ.
- $MDS(\sigma, n)$: the *maximum* number of *distinct* Lyndon subsequences in a string of length n over an alphabet of size σ.
- $ETS(\sigma, n)$: the *expected total* number of Lyndon subsequences in a string of length n over an alphabet of size σ.
- $EDS(\sigma, n)$: the *expected* number of *distinct* Lyndon subsequences in a string of length n over an alphabet of size σ.

Among these, we present the exact values for $MTS(\sigma, n)$, $ETS(\sigma, n)$, and $EDS(\sigma, n)$. Our results are summarized in Table 2. Although the main ideas of our proofs are analogous to the results for substrings, there exist differences based on properties of substrings and subsequences.

Number of Lyndon Subsequences in a String	
Maximum Total $MTS(\sigma, n)$	$2^n - (p + \sigma)2^m + n + \sigma - 1$ [this work]
Maximum Distinct $MDS(\sigma, n)$	open
Expected Total $ETS(\sigma, n)$	$\sum_{m=1}^{n} \left[L(\sigma, m) \binom{n}{m} \sigma^{n-m} \right] \sigma^{-n}$ [this work]
Expected Distinct $EDS(\sigma, n)$	$\sum_{m=1}^{n} \left[L(\sigma, m) \sum_{k=m}^{n} \binom{n}{k} (\sigma - 1)^{n-k} \right] \sigma^{-n}$ [this work]

Table 2. The numbers of Lyndon subsequences in a string of length n over an alphabet of size σ, where $n = m\sigma + p$ with $0 \le p < \sigma$ for $MTS(\sigma, n)$.

In the future work, we hope to determine the exact value for $MDS(\sigma, n)$.

2 Preliminaries

2.1 Strings

Let $\Sigma = \{a_1, \ldots, a_\sigma\}$ be an ordered *alphabet* of size σ such that $a_1 < \ldots < a_\sigma$. An element of Σ^* is called a *string*. The length of a string w is denoted by $|w|$. The empty string ε is a string of length 0. Let Σ^+ be the set of non-empty strings, i.e., $\Sigma^+ = \Sigma^* - \{\varepsilon\}$. The i-th character of a string w is denoted by $w[i]$, where $1 \le i \le |w|$. For a string w and two integers $1 \le i \le j \le |w|$, let $w[i..j]$ denote the substring of w that begins at position i and ends at position j. For convenience, let $w[i..j] = \varepsilon$ when $i > j$. A string x is said to be a subsequence of a string w if there exists a set of positions $\{i_1, \ldots, i_{|x|}\}$ $(1 \le i_1 < \ldots < i_{|x|} \le |w|)$ such that $x = w[i_1] \cdots w[i_{|x|}]$. We say that a subsequence x occurs at $\{i_1, \ldots, i_{|x|}\}$ $(1 \le i_1 < \ldots < i_{|x|} \le |w|)$ if $x = w[i_1] \cdots w[i_{|x|}]$.

2.2 Lyndon words

A string $x = uv$ is said to be a *conjugate* of another string y if $y = vu$. A string w is called a *Lyndon word* if it is the lexicographically smallest among all of its conjugates. Equivalently, a string w is said to be a Lyndon word, if w is lexicographically smaller than all of its non-empty proper suffixes.

Let μ be the *Möbius function* on the set of positive integers defined as follows.

$$\mu(n) = \begin{cases} 1 & (n = 1) \\ 0 & (\text{if } n \text{ is divisible by a square}) \\ (-1)^k & (\text{if } n \text{ is the product of } k \text{ distinct primes}) \end{cases}$$

It is known that the number $L(\sigma, n)$ of Lyndon words of length n over an alphabet of size σ can be represented as

$$L(\sigma, n) = \frac{1}{n} \sum_{d|n} \mu\left(\frac{n}{d}\right) \sigma^d,$$

where $d|n$ is the set of divisors d of n [6].

3 Maximum total number of Lyndon subsequences

Let $MTS(\sigma, n)$ be the maximum total number of Lyndon subsequences in a string of length n over an alphabet Σ of size σ. In this section, we determine $MTS(\sigma, n)$.

Theorem 1. *For any σ and n such that $\sigma < n$,*

$$MTS(\sigma, n) = 2^n - (p + \sigma)2^m + n + \sigma - 1$$

where $n = m\sigma + p$ $(0 \le p < \sigma)$. Moreover, the number of strings that contain $MTS(\sigma, n)$ Lyndon subsequences is $\binom{\sigma}{p}$, and the following string w is one of such strings;

$$w = {a_1}^m \cdots {a_{\sigma-p}}^m {a_{\sigma-p+1}}^{m+1} \cdots {a_\sigma}^{m+1}.$$

Proof. Consider a string w of the form

$$w = {a_1}^{k_1}{a_2}^{k_2}\cdots{a_\sigma}^{k_\sigma}$$

where $\sum_{i=1}^{\sigma} k_i = n$ and $k_i \geq 0$ for any i. For any subsequence x of w, x is a Lyndon word if x is not a unary string of length at least 2. It is easy to see that this form is a necessary condition for the maximum number ($\because$ there exist several non-Lyndon subsequences if $w[i] > w[j]$ for some $i < j$). Hence, the number of Lyndon subsequences of w can be represented as

$$\begin{aligned}(2^n - 1) - \sum_{i=1}^{\sigma}(2^{k_i} - 1 - k_i) &= 2^n - 1 - \sum_{i=1}^{\sigma} 2^{k_i} + \sum_{i=1}^{\sigma} k_i + \sigma \\ &= 2^n - 1 - \sum_{i=1}^{\sigma} 2^{k_i} + n + \sigma.\end{aligned}$$

This formula is maximized when $\sum_{i=1}^{\sigma} 2^{k_i}$ is minimized. It is known that

$$2^a + 2^b > 2^{a-1} + 2^{b+1}$$

holds for any integer a, b such that $a \geq b + 2$. From this fact, $\sum_{i=1}^{\sigma} 2^{k_i}$ is minimized when the difference of k_i and k_j is less than or equal to 1 for any i, j. Thus, if we choose p k_i's as $m+1$, and set m for other $(\sigma - p)$ k_i's where $n = m\sigma + p$ $(0 \leq p < \sigma)$, then $\sum_{i=1}^{\sigma} 2^{k_i}$ is minimized. Hence,

$$\begin{aligned}\min(2^n - 1 - \sum_{i=1}^{\sigma} 2^{k_i} + n + \sigma) &= 2^n - 1 - p \cdot 2^{m+1} - (\sigma - p)2^m + n + \sigma \\ &= 2^n - (p + \sigma)2^m + n + \sigma - 1\end{aligned}$$

Moreover, one of such strings is

$${a_1}^m \cdots {a_{\sigma-p}}^m {a_{\sigma-p+1}}^{m+1} \cdots {a_\sigma}^{m+1}.$$

Therefore, this theorem holds. □

We can apply the above strategy to the version of substrings. Namely, we can also obtain the following result.

Corollary 2. *Let $MTF(\sigma, n)$ be the maximum total number of Lyndon substrings in a string of length n over an alphabet of size σ. For any σ and n such that $\sigma < n$,*

$$MTF(\sigma, n) = \binom{n}{2} - (\sigma - p)\binom{m+1}{2} - p\binom{m+2}{2} + n$$

where $n = m\sigma + p$ $(0 \leq p < \sigma)$. Moreover, the number of strings that contain $MTF(\sigma, n)$ Lyndon subsequences is $\binom{\sigma}{p}$, and the following string w is one of such strings;

$$w = {a_1}^m \cdots {a_{\sigma-p}}^m {a_{\sigma-p+1}}^{m+1} \cdots {a_\sigma}^{m+1}.$$

Proof. Consider a string w of the form

$$w = {a_1}^{k_1}{a_2}^{k_2}\cdots{a_\sigma}^{k_\sigma}$$

where $\sum_{i=1}^{\sigma} k_i = n$ and $k_i \geq 0$ for any i. In a similar way to the above discussion, the number of Lyndon substrings of w can be represented as

$$\binom{n+1}{2} - \sum_{i=1}^{\sigma} \left[\binom{k_i+1}{2} - k_i \right] = \binom{n+1}{2} - \sum_{i=1}^{\sigma} \binom{k_i+1}{2} + n.$$

We can use the following inequation that holds for any a, b such that $a \geq b+2$;

$$\binom{a}{2} + \binom{b}{2} > \binom{a-1}{2} + \binom{b+1}{2}.$$

Then,

$$\min \left[\binom{n+1}{2} - \sum_{i=1}^{\sigma} \binom{k_i+1}{2} + n \right] = \binom{n}{2} - (\sigma - p)\binom{m+1}{2} - p\binom{m+2}{2} + n$$

holds. □

Finally, we give exact values $MTS(\sigma, n)$ for several conditions in Table 3.

n	$MTS(2,n)$	$MTS(5,n)$	$MTS(10,n)$
1	1	1	1
2	3	3	3
3	6	7	7
4	13	15	15
5	26	31	31
6	55	62	63
7	122	125	127
8	233	252	255
9	474	507	511
10	971	1018	1023
11	1964	2039	2046
12	3981	4084	4093
13	8014	8177	8188
14	16143	16366	16379
15	32400	32747	32762

Table 3. Values $MTS(\sigma, n)$ for $\sigma = 2, 5, 10$, $n = 1, 2, \cdots, 15$.

4 Expected total number of Lyndon subsequences

Let $TS(\sigma, n)$ be the total number of Lyndon subsequences in all strings of length n over an alphabet Σ of size σ. In this section, we determine the expected total number $ETS(\sigma, n)$ of Lyndon subsequences in a string of length n over an alphabet Σ of size σ, namely, $ETS(\sigma, n) = TS(\sigma, n)/\sigma^n$.

Theorem 3. *For any σ and n such that $\sigma < n$,*

$$TS(\sigma, n) = \sum_{m=1}^{n} \left[L(\sigma, m) \binom{n}{m} \sigma^{n-m} \right].$$

Moreover, $ETS(\sigma, n) = TS(\sigma, n)/\sigma^n$.

Proof. Let $Occ(w, x)$ be the number of occurrences of subsequence x in w, and $L(\sigma, n)$ the set of Lyndon words of length less than or equal to n over an alphabet of size σ. By a simple observation, $TS(\sigma, n)$ can be written as

$$TS(\sigma, n) = \sum_{x \in \mathcal{L}(\sigma,n)} \sum_{w \in \Sigma^n} Occ(w, x).$$

Firstly, we consider $\sum_{w \in \Sigma^n} Occ(w, x)$ for a Lyndon word x of length m. Let $\{i_1, \ldots, i_m\}$ be a set of m positions in a string of length n where $1 \le i_1 < \ldots < i_m \le n$. The number of strings that contain x as a subsequence at $\{i_1, \ldots, i_m\}$ is σ^{n-m}. In addition, the number of combinations of m positions is $\binom{n}{m}$. Hence, $\sum_{w \in \Sigma^n} Occ(w, x) = \binom{n}{m}\sigma^{n-m}$. This implies that

$$TS(\sigma, n) = \sum_{m=1}^{n} \left[L(\sigma, m) \binom{n}{m} \sigma^{n-m} \right].$$

Finally, since the number of strings of length n over an alphabet of size σ is σ^n, $ETS(\sigma, n) = TS(\sigma, n)/\sigma^n$. Therefore, this theorem holds. □

Finally, we give exact values $TS(\sigma, n), ETS(\sigma, n)$ for several conditions in Table 4.

n	$TS(2,n)$	$ETS(2,n)$	$TS(5,n)$	$ETS(5,n)$
1	2	1.00	5	1.00
2	9	2.25	60	2.40
3	32	4.00	565	4.52
4	107	6.69	4950	7.92
5	356	11.13	42499	13.60
6	1205	18.83	365050	23.36
7	4176	32.63	3163435	40.49
8	14798	57.80	27731650	70.99
9	53396	104.29	245950375	125.93
10	195323	190.75	2204719998	225.76

Table 4. Values $TS(\sigma, n), ETS(\sigma, n)$ for $\sigma = 2, 5, n = 1, 2, \cdots, 10$.

5 Expected number of distinct Lyndon subsequences

Let $TDS(\sigma, n)$ be the total number of distinct Lyndon subsequences in all strings of length n over an alphabet Σ of size σ. In this section, we determine the expected number $EDS(\sigma, n)$ of distinct Lyndon subsequences in a string of length n over an alphabet Σ of size σ, namely, $EDS(\sigma, n) = TDS(\sigma, n)/\sigma^n$.

Theorem 4. *For any σ and n such that $\sigma < n$,*

$$TDS(\sigma, n) = \sum_{m=1}^{n} \left[L(\sigma, m) \sum_{k=m}^{n} \binom{n}{k} (\sigma - 1)^{n-k} \right].$$

Moreover, $EDS(\sigma, n) = TDS(\sigma, n)/\sigma^n$.

To prove this theorem, we introduce the following lemmas.

Lemma 5. *For any $x_1, x_2 \in \Sigma^m$ and m, n $(m \leq n)$, the number of strings in Σ^n which contain x_1 as a subsequence is equal to the number of strings in Σ^n which contain x_2 as a subsequence.*

Proof (of Lemma 5). Let $C(n, \Sigma, x)$ be the number of strings in Σ^n which contain a string x as a subsequence. We prove $C(n, \Sigma, x_1) = C(n, \Sigma, x_2)$ for any $x_1, x_2 \in \Sigma^m$ by induction on the length m.

Suppose that $m = 1$. It is clear that the set of strings which contain $x \in \Sigma$ is $\Sigma^n - (\Sigma - \{x\})^n$, and $C(n, \Sigma, x) = \sigma^n - (\sigma - 1)^n$. Thus, $C(n, \Sigma, x_1) = C(n, \Sigma, x_2)$ for any x_1, x_2 if $|x_1| = |x_2| = 1$.

Suppose that the statement holds for some $k \geq 1$. We prove $C(n, \Sigma, x_1) = C(n, \Sigma, x_2)$ for any $x_1, x_2 \in \Sigma^{k+1}$ by induction on n. If $n = k+1$, then $C(n, \Sigma, x_1) = C(n, \Sigma, x_2) = 1$. Assume that the statement holds for some $\ell \geq k+1$. Let $x = yc$ be a string of length $k+1$ such that $y \in \Sigma^k, c \in \Sigma$. Each string w of length $\ell + 1$ which contains x as a subsequence satisfies either

- $w[1..\ell]$ contains x as a subsequence, or
- $w[1..\ell]$ does not contain x as a subsequence.

The number of strings w in the first case is $\sigma \cdot C(j, \Sigma, yc)$. On the other hand, the number of strings w in the second case is $C(\ell, \Sigma, y) - C(\ell, \Sigma, yc)$. Hence, $C(\ell + 1, \Sigma, x) = \sigma C(\ell, \Sigma, yc) + C(\ell, \Sigma, y) - C(\ell, \Sigma, yc)$. Let $x_1 = y_1 c_1$ and $x_2 = y_2 c_2$ be strings of length $k+1$. By an induction hypothesis, $C(\ell, \Sigma, y_1 c_1) = C(\ell, \Sigma, y_2 c_2)$ and $C(\ell, \Sigma, y_1) = C(\ell, \Sigma, y_2)$ hold. Thus, $C(\ell + 1, \Sigma, x_1) = C(\ell + 1, \Sigma, x_2)$ also holds.

Therefore, this lemma holds. □

Lemma 6. *For any string x of length $m \leq n$,*

$$C(n, \Sigma, x) = \sum_{k=m}^{n} \binom{n}{k} (\sigma - 1)^{n-k}.$$

Proof (of Lemma 6). For any character c, it is clear that the number of strings that contain c exactly k times is $\binom{n}{k}(\sigma - 1)^{n-k}$. By Lemma 5,

$$C(n, \Sigma, x) = C(n, \Sigma, c^m) = \sum_{k=m}^{n} \binom{n}{k} (\sigma - 1)^{n-k}.$$

Hence, this lemma holds. □

Then, we can obtain Theorem 4 as follows.

Proof (of Theorem 4). Thanks to Lemma 6, the number of strings of length n which contain a Lyndon word of length m is also $\sum_{k=m}^{n} \binom{n}{k}(\sigma - 1)^{n-k}$. Since the number of Lyndon words of length m over an alphabet of size σ is $L(\sigma, m)$,

$$TDS(\sigma, n) = \sum_{m=1}^{n} \left[L(\sigma, m) \sum_{k=m}^{n} \binom{n}{k} (\sigma - 1)^{n-k} \right].$$

Finally, since the number of strings of length n over an alphabet of size σ is σ^n, $EDS(\sigma, n) = TDS(\sigma, n)/\sigma^n$. Therefore, Theorem 4 holds. □

We give exact values $EDS(\sigma, n)$ for several conditions in Table 5.

n	$EDS(2,n)$	$EDS(5,n)$
1	1.00	1.00
2	1.75	2.20
3	2.50	3.80
4	3.38	6.09
5	4.50	9.51
6	6.00	14.80
7	8.03	23.12
8	10.81	36.43
9	14.63	57.95
10	19.93	93.08
15	100.57	1121.29
20	559.42	15444.90

Table 5. Values $EDS(\sigma, n)$ for $\sigma = 2, 5, n = 1, \ldots, 10, 15, 20$.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP18K18002 (YN), JP21K17705 (YN), JP18H04098 (MT), JST ACT-X Grant Number JPMJAX200K (YN), and by JST PRESTO Grant Number JPMJPR1922 (SI).

References

1. M. A. Babenko, P. Gawrychowski, T. Kociumaka, I. I. Kolesnichenko, and T. Starikovskaya: *Computing minimal and maximal suffixes of a substring.* Theor. Comput. Sci., 638 2016, pp. 112–121.
2. H. Bannai, T. I, S. Inenaga, Y. Nakashima, M. Takeda, and K. Tsuruta: *The "runs" theorem.* SIAM J. Comput., 46(5) 2017, pp. 1501–1514.
3. M. Crochemore and D. Perrin: *Two-way string matching.* J. ACM, 38(3) 1991, pp. 651–675.
4. S. Giuliani, S. Inenaga, Z. Lipták, N. Prezza, M. Sciortino, and A. Toffanello: *Novel results on the number of runs of the Burrows-Wheeler-transform*, in SOFSEM 2021, vol. 12607 of Lecture Notes in Computer Science, Springer, 2021, pp. 249–262.
5. A. Glen, J. Simpson, and W. F. Smyth: *Counting Lyndon factors.* The Electronic Journal of Combinatorics, 24 2017, p. P3.28.
6. M. Lothaire: *Combinatorics on Words*, Addison-Wesley, 1983.
7. M. Mucha: *Lyndon words and short superstrings*, in Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013, SIAM, 2013, pp. 958–972.
8. R. Sugahara, Y. Nakashima, S. Inenaga, H. Bannai, and M. Takeda: *Computing runs on a trie*, in CPM 2019, vol. 128 of LIPIcs, 2019, pp. 23:1–23:11.
9. K. Tsuruta, D. Köppl, Y. Nakashima, S. Inenaga, H. Bannai, and M. Takeda: *Grammar-compressed self-index with Lyndon words.* IPSJ Transactions on Mathematical Modeling and its Applications (TOM), 13(2) 2020, pp. 84–92.