Counting Lyndon Subsequences

Ryo Hirakawa, Yuto Nakashima,
Shunsuke Inenaga, and Masayuki Takeda
Kyushu University, Japan

Background

- Lyndon factors enjoy a rich class of algorithmic and stringology applications.
  - counting and finding the maximal repetitions in a string [Bannai et al., 2017]

- One of the mathematical interests for Lyndon words is counting the number of Lyndon words in a string.
  - Glen et al. (2017) studied # of Lyndon factors in strings.
  - We study # of Lyndon subsequences in strings.

Lyndon word [Lyndon, 1954]

<u>Definition</u>
A word w is a **Lyndon word** if w is lexicographically smaller than all of its non-empty proper suffixes.

- w = aabab is a Lyndon word.

$$\begin{array}{rr} w = & \text{aabab} \\ w \prec & \text{abab} \\ w \prec & \text{bab} \\ w \prec & \text{ab} \\ w \prec & \text{b} \end{array}\left.\vphantom{\begin{array}{r}a\\a\\a\\a\end{array}}\right\} w\text{'s non-empty proper suffixes}$$

Lyndon factor/subsequence

<u>Definition</u>

- A factor f of string w is called a **Lyndon factor** if f is a Lyndon word.
- A subsequence s of string w is called a **Lyndon subsequence** if s is a Lyndon word.

- $f =$ aab is a Lyndon factor of w = b**aab**ab.
- $s =$ aabb is a Lyndon subsequence of w = b**aab**a**b**.

Maximum number of distinct Lyndon factors

- Glen et al. counted the maximum number of distinct Lyndon factors in Σ^n.
- w = bcabaababa contains 7 distinct Lyndon factors.

w = b c a b a a b a b a

1: **b** c a b a a b a b a

2: b **c** a b a a b a b a

3: b c **a** b a a b a b a

4: **b c** a b a a b a b a

5: b c **a b** a a b a b a

6: b c a b **a a b** a b a

7: b c a b **a a b a b** a

Maximum number of distinct Lyndon factors

Theorem 1 [Glen et al., 2017]

$$MDF(\sigma, n) \coloneqq \max_{w \in \Sigma^n} (\text{ the number of distinct Lyndon factors in } w \text{ })$$
$$= \binom{n+1}{2} - (\sigma - 1)\binom{q+1}{2} - r\binom{q+2}{2} + \sigma$$

One of the strings that gives this value is $w = a_1^{q+1} \cdots a_r^{q+1}\ a_{r+1}^{q} \cdots a_\sigma^{q}$

where $a_1 \prec a_2 \prec \cdots \prec a_\sigma, n = q\sigma + r\ (0 \le r < \sigma)$.

- For example if $\Sigma = \{$a,b,c$\}$, $n = 10$ then, $MDF(3{,}10) = 36$. One of such strings is w = aaaabbbccc.

Total number of Lyndon factor occ.

- Glen et al. counted the total number of Lyndon factors appearing in all strings in Σ^n.
- $\Sigma = \{\mathrm{a,b}\}$, $n = 3$

$\{\mathrm{a,b}\}^3$	aaa	aab	aba	abb	baa	bab	bba	bbb	
Lyndon factors	**a**aa a**a**a aa**a**	**a**ab a**a**b aa**b** a**ab** **aab**	**a**ba a**b**a ab**a** **ab**a	**a**bb a**b**b ab**b** **ab**b **abb**	**b**aa b**a**a ba**a**	**b**ab b**a**b ba**b** b**ab**	**b**ba b**b**a bb**a**	**b**bb b**b**b bb**b**	
number	3	5	4	5	3	4	3	3	30

This implies that the expected number of Lyndon factors in a string in $\{\mathrm{a,b}\}^3$ is 30/8.

Total number of Lyndon factor occ.

Theorem 2 [Glen et al., 2017]

$TF(\sigma, n) \coloneqq \sum_{w \in \Sigma^n} ($ the total number of Lyndon factor occ. in w $)$

$= \sum_{m=1}^{n} (L(\sigma, m)\ (n-m+1)\ \sigma^{n-m})$

Lemma 1 [Lothaire, 1997]

Let $L(\sigma, m)$ be the number of Lyndon words in Σ^m. Then,

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

μ is the Möbius function.

Total number of Lyndon factor occ.

<u>Theorem 2</u> [Glen et al., 2017]

$TF(\sigma, n) \coloneqq \sum_{w \in \Sigma^n}($ the total number of Lyndon factor occ. in w $)$

$= \sum_{m=1}^{n} (L(\sigma, m)\ (n - m + 1)\ \sigma^{n-m})$

- The expected total number of Lyndon factors $ETF(\sigma, n)$ is given as follows.

$$ETF(\sigma, n) = \frac{TF(\sigma, n)}{\sigma^n}$$

Total number of distinct Lyndon factors

- Glen et al. counted the total number of distinct Lyndon factors in all strings in Σ^n.
- $\Sigma = \{\text{a,b}\}$, $n = 3$

$\{\text{a,b}\}^3$	aaa	aab	aba	abb	baa	bab	bba	bbb	
distinct Lyndon factors	**a**aa	**a**ab aa**b** a**ab** **aab**	**a**ba a**b**a **ab**a	**a**bb a**b**b **ab**b **abb**	**b**aa b**a**a	**b**ab b**a**b b**ab**	**b**ba bb**a**	**b**bb	
number	1	4	3	4	2	3	2	1	20

This implies that the <u>expected number</u> of distinct Lyndon factors in a string in $\{\text{a,b}\}^3$ is 20/8.

Total number of distinct Lyndon factors

Theorem 3 [Glen et al., 2017]

$DF(\sigma, n) := \sum_{w \in \Sigma^n}($ the number of distinct Lyndon factors in w $)$

$$= \sum_{m=1}^{n} \left(L(\sigma, m) \sum_{k=1}^{\lfloor n/m \rfloor} (-1)^{k+1} \binom{n-km+k}{k} \sigma^{n-km} \right)$$

$L(\sigma, m)$ is the number of Lyndon words in Σ^m.

- The expected number of distinct Lyndon factors $EDF(\sigma, n)$ is given as follows.

$$EDF(\sigma, n) = \frac{DF(\sigma, n)}{\sigma^n}$$

Previous work

- Glen et al. showed three theorems in table below.

	factor	subsequence
max. distinct	$\binom{n+1}{2} - (\sigma - 1)\binom{q+1}{2} - r\binom{q+2}{2} + \sigma$	
max. total		
expected total	$\sum_{m=1}^{n} \left(L(\sigma, m)\,(n - m + 1)\,\sigma^{-m} \right)$	
expected distinct	$\sum_{m=1}^{n} \left(L(\sigma, m) \sum_{k=1}^{\lfloor n/m \rfloor} (-1)^{k+1} \binom{n - km + k}{k} \sigma^{-km} \right)$	

n: length of string , σ : alphabet size, $L(\sigma, m)$: # of Lyndon words in Σ^m
$n = q\sigma + r \ (0 \leq r < \sigma)$

Our work

- We obtained four theorems below.

	factor	subsequence
max. distinct	$\binom{n+1}{2} - (\sigma-1)\binom{q+1}{2} - r\binom{q+2}{2} + \sigma$	
max. total	$\binom{n+1}{2} - (\sigma-1)\binom{q+1}{2} - r\binom{q+2}{2} + n$	$2^n - (r+\sigma)2^q + n + \sigma - 1$
expected total	$\sum_{m=1}^{n} \left(L(\sigma,m)\,(n-m+1)\,\sigma^{-m}\right)$	$\sum_{m=1}^{n} \left(L(\sigma,m)\binom{n}{m}\sigma^{-m}\right)$
expected distinct	$\sum_{m=1}^{n} \left(L(\sigma,m) \sum_{k=1}^{\lfloor n/m \rfloor} (-1)^{k+1} \binom{n-km+k}{k} \sigma^{-km}\right)$	$\sum_{m=1}^{n} \left(L(\sigma,m) \sum_{k=m}^{n} \binom{n}{k} (\sigma-1)^{n-k}\right) \sigma^{-n}$

n: length of string , σ : alphabet size, $L(\sigma,m)$: # of Lyndon words in Σ^m

$n = q\sigma + r \ (0 \le r < \sigma)$

Our work

- We obtained four theorems below.

	factor	subsequence
max. distinct	$\binom{n+1}{2} - (\sigma - 1)\binom{q+1}{2} - r\binom{q+2}{2} + \sigma$	
max. total	$\binom{n+1}{2} - (\sigma - 1)\binom{q+1}{2} - r\binom{q+2}{2} + n$	$2^n - (r+\sigma)2^q + n + \sigma - 1$
expected total	$\sum_{m=1}^{n} \left(L(\sigma,m)\,(n-m+1)\,\sigma^{-m}\right)$	$\sum_{m=1}^{n} \left(L(\sigma,m)\binom{n}{m}\sigma^{-m}\right)$
expected distinct	$\sum_{m=1}^{n} \left(L(\sigma,m) \sum_{k=1}^{\lfloor n/m \rfloor} (-1)^{k+1} \binom{n-km+k}{k} \sigma^{-km}\right)$	$\sum_{m=1}^{n} \left(L(\sigma,m) \sum_{k=m}^{n} \binom{n}{k} (\sigma-1)^{n-k}\right)\sigma^{-n}$

n: length of string , σ : alphabet size, $L(\sigma, m)$: # of Lyndon words in Σ^m

$n = q\sigma + r \ (0 \leq r < \sigma)$

Maximum total number of Lyndon subsequence occ.

<u>Theorem 4</u>

$MTS(\sigma, n) \coloneqq \max_{w \in \Sigma^n}(\text{the total number of Lyndon subsequence occ. in } w)$
$= 2^n - (r+\sigma)2^q + n + \sigma - 1$

One of the strings that gives this value is $w = a_1^{q+1} \cdots a_r^{q+1} \, a_{r+1}^q \cdots a_\sigma^q$
where $a_1 \prec a_2 \prec \cdots \prec a_\sigma, n = q\sigma + r \; (0 \le r < \sigma)$.

- For example if $\Sigma = \{$a,b,c$\}$, $n = 10$ then, $MDF(3,10) = 1004$.
  One of such strings is

w = a a a a b b b c c c .

Maximum total number of Lyndon subsequence occ.

Theorem 4

$MTS(\sigma, n) \coloneqq \max_{w \in \Sigma^n}$(the total number of Lyndon subsequence occ. in w)

$$= 2^n - (r + \sigma)2^q + n + \sigma - 1$$

One of the strings that gives this value is $w = a_1^{q+1} \cdots a_r^{q+1}\ a_{r+1}^q \cdots a_\sigma^q$

where $a_1 \prec a_2 \prec \cdots \prec a_\sigma$, $n = q\sigma + r\ (0 \leq r < \sigma)$.

- For example if $\Sigma = \{$a,b,c$\}$, $n = 10$ then, $MDF(3,10) = 1004$.
  One of such strings is

w = a a a a b b b c c c .

If w is a lexicographically non-decreasing string,
then any subsequence of w that contains
at least two distinct symbols is a Lyndon word.

Maximum total number of Lyndon subsequence occ.

Theorem 4

$$MTS(\sigma, n) \coloneqq \max_{w \in \Sigma^n}(\text{the total number of Lyndon subsequence occ. in } w)$$
$$= 2^n - (r+\sigma)2^q + n + \sigma - 1$$

One of the strings that gives this value is $w = a_1^{q+1} \cdots a_r^{q+1} \ a_{r+1}^q \cdots a_\sigma^q$ where $a_1 \prec a_2 \prec \cdots \prec a_\sigma, n = q\sigma + r$ $(0 \leq r < \sigma)$.

- For example if $\Sigma = \{$a,b,c$\}$, $n = 10$ then, $MDF(3,10) = 1004$. One of such strings is

not Lyndon

$w =$ a a a a b b b c c c .

String w is optimal to minimize the number of subsequences of the form a^m $(m \geq 2)$ which is not Lyndon.

Exact values of $MTS(\sigma, n)$

n \ σ	2	5	10
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

$MTS(\sigma, n)$

Total number of Lyndon subsequence occurrences

<u>Theorem 5</u>

$$TS(\sigma,n) := \sum_{w\in\Sigma^n} (\text{ the total number of Lyndon subsequence occ. in } w\ )$$

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

$L(\sigma, m)$ is the number of Lyndon words in Σ^m.

- $\Sigma = \{\text{a,b}\}$, $n = 3$

$\{\text{a,b}\}^3$	aaa	aab	aba	abb	baa	bab	bba	bbb	
occurrences of Lyndon subsequences	**a**aa a**a**a aa**a**	**a**ab a**a**b aa**b** a**ab** **a**a**b** **aab**	**a**ba a**b**a ab**a** **ab**a	**a**bb a**b**b ab**b** **ab**b **a**b**b** **abb**	**b**aa b**a**a ba**a**	**b**ab b**a**b ba**b** b**ab**	**b**ba b**b**a bb**a**	**b**bb b**b**b bb**b**	$TS(2,3)$
the number	3	6	4	6	3	4	3	3	32

Total number of Lyndon subsequence occurrences

Theorem 5

$$TS(\sigma, n) := \sum_{w \in \Sigma^n} (\text{ the total number of Lyndon subsequence occ. in } w\ )$$

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

Expected total number $ETS(2,3)$
of Lyndon subsequences
in a string is 32/8.

- $\Sigma = \{$a,b$\}$, $n = 3$

$\{$a,b$\}^3$	aaa	aab	aba	abb	baa	bab	bba	bbb	
occurrences of Lyndon subsequences	**a**aa a**a**a aa**a**	**a**ab a**a**b aa**b** a**ab** **a**a**b** **aab**	**a**ba a**b**a ab**a** **ab**a	**a**bb a**b**b ab**b** **ab**b **a**b**b** **abb**	**b**aa b**a**a ba**a**	**b**ab b**a**b ba**b** b**ab**	**b**ba b**b**a bb**a**	**b**bb b**b**b bb**b**	$TS(2,3)$
the number	3	6	4	6	3	4	3	3	32

Total number of Lyndon subsequence occurrences

<u>Theorem 5</u>

$$TS(\sigma, n) := \sum_{w \in \Sigma^n} (\text{ the total number of Lyndon subsequence occ. in } w \text{ })$$

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

$L(\sigma, m)$ is the number of Lyndon words in Σ^m.

- Consider how many times a Lyndon word x of length $m \leq n$ occurs as a subsequence of strings in Σ^n.

Total number of Lyndon subsequence occurrences

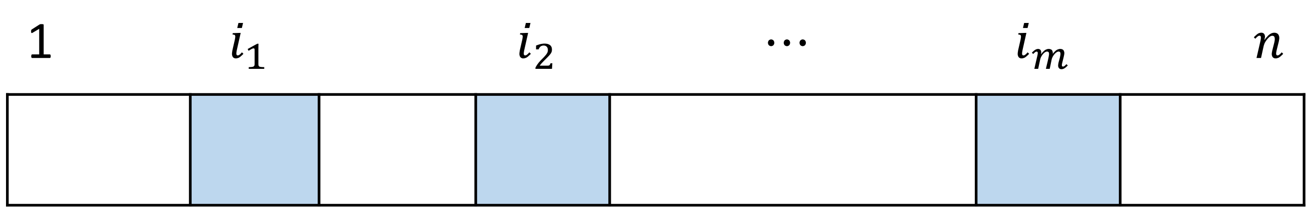
<u>Theorem 5</u>

$$TS(\sigma, n) := \sum_{w \in \Sigma^n} (\text{ the total number of Lyndon subsequence occ. in } w\ )$$

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

$L(\sigma, m)$ is the number of Lyndon words in Σ^m.

- Let $\{i_1, i_2, \ldots, i_m\}$ be a set of m positions of string of length n.



Total number of Lyndon subsequence occurrences

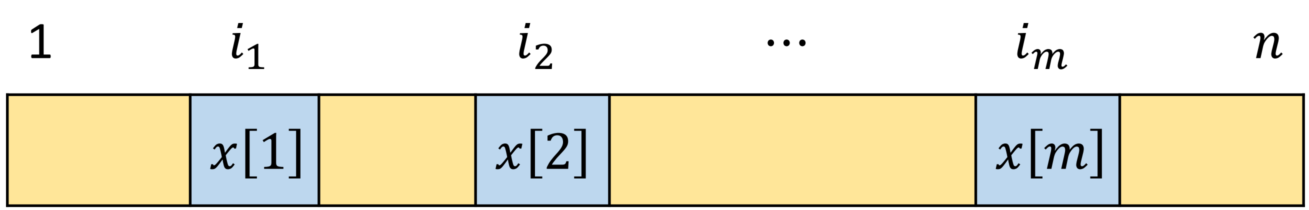
<u>Theorem 5</u>

$TS(\sigma, n) := \sum_{w \in \Sigma^n}$ (the total number of Lyndon subsequence occ. in w)

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

$L(\sigma, m)$ is the number of Lyndon words in Σ^m.

- Let $\{i_1, i_2, \ldots, i_m\}$ be a set of m positions of string of length n.
- The number of strings containing x at these positions is σ^{n-m}.



Total number of Lyndon subsequence occurrences

<u>Theorem 5</u>

$$TS(\sigma, n) := \sum_{w \in \Sigma^n} (\text{ the total number of Lyndon subsequence occ. in } w\text{ })$$

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

$L(\sigma, m)$ is the number of Lyndon words in Σ^m.

- Let $\{i_1, i_2, \ldots, i_m\}$ be a set of m positions of string of length n.
- The number of strings containing x at these positions is σ^{n-m}.
- The number of combinations of m positions is $\binom{n}{m}$.

1	i_1		i_2	$\cdots$	i_m	n
	$x[1]$		$x[2]$		$x[m]$	

Total number of Lyndon subsequence occurrences

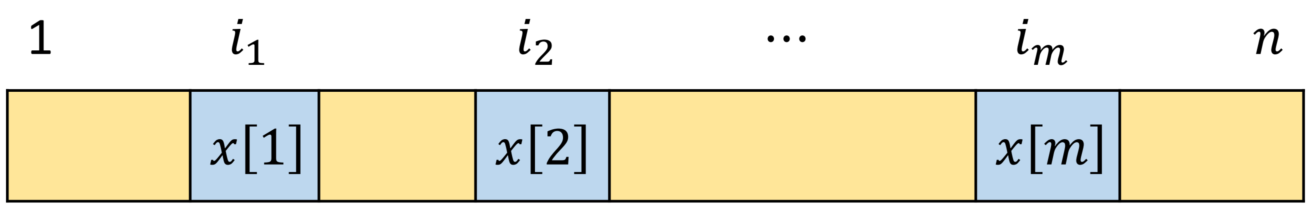
<u>Theorem 5</u>

$$TS(\sigma, n) \coloneqq \sum_{w \in \Sigma^n} (\text{ the total number of Lyndon subsequence occ. in } w\ )$$

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

$L(\sigma, m)$ is the number of Lyndon words in Σ^m.

- Let $\{i_1, i_2, \ldots, i_m\}$ be a set of m positions of string of length n.
- The number of strings containing x at these positions is σ^{n-m}.
- The number of combinations of m positions is $\binom{n}{m}$.
- The number of Lyndon words of length m is $L(\sigma, m)$.



Total number of Lyndon subsequence occurrences

<u>Theorem 5</u>

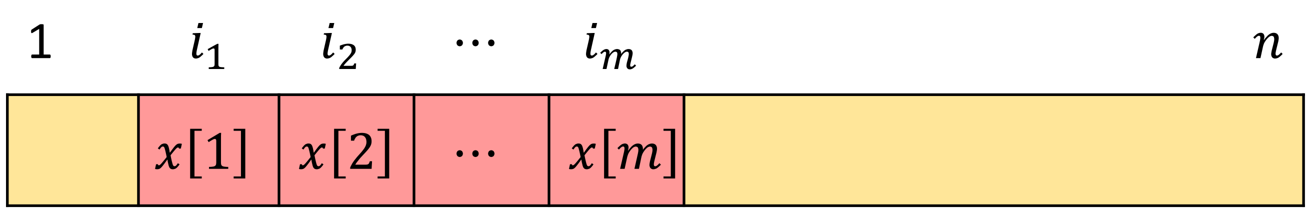
$$TS(\sigma, n) \coloneqq \sum_{w \in \Sigma^n} (\text{ the total number of Lyndon subsequence occ. in } w\ )$$
$$= \sum_{m=1}^{n} \left(L(\sigma, m) \binom{n}{m} \sigma^{n-m} \right)$$

$L(\sigma, m)$ is the number of Lyndon words in Σ^m.

- If m positions have to be contiguous, there are $(n - m + 1)$ ways to choose such positions.

<u>Theorem 2</u> [Glen et al., 2017]

$$TF(\sigma, n) \coloneqq \sum_{w \in \Sigma^n} (\text{ the total number of Lyndon factor occ. in } w\ )$$
$$= \sum_{m=1}^{n} (L(\sigma, m)\, (n - m + 1)\, \sigma^{n-m})$$



Exact values of $ETF(\sigma, n)$ and $ETS(\sigma, n)$

n \ σ	2	5
1	1.00	1.00
2	2.25	2.40
3	3.75	4.12
4	5.43	6.08
5	7.31	8.24
6	9.32	10.56
7	11.48	13.03
8	13.76	15.62
9	16.14	18.33
10	18.62	21.13

$ETF(\sigma, n)$

(expected total #of Lyndon **factors**)

n \ σ	2	5
1	1.00	1.00
2	2.25	2.40
3	4.00	4.52
4	6.69	7.92
5	11.13	13.60
6	18.83	23.36
7	32.63	40.49
8	57.80	70.99
9	104.29	125.93
10	190.75	225.76

$ETS(\sigma, n)$

(expected total # of Lyndon **subsequences**)

Total number of distinct Lyndon subsequences

Theorem 6

$$TDS(\sigma, n) := \sum_{w \in \Sigma^n} (\text{ the number of distinct Lyndon subsequences in } w \ )$$

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

$L(\sigma, m)$ is the number of Lyndon words in Σ^m.

- $\Sigma = \{\text{a,b}\}$, $n = 3$

$\{\text{a,b}\}^3$	aaa	aab	aba	abb	baa	bab	bba	bbb	
distinct Lyndon subsequences	**a**aa	**a**ab aa**b** **a**a**b** **aab**	**a**ba a**b**a **ab**a	**a**bb a**b**b **a**b**b** **abb**	**b**aa b**a**a	**b**ab b**a**b b**ab**	**b**ba bb**a**	**b**bb	
the number	1	4	3	4	2	3	2	1	20

$TDS(2,3)$

Total number of distinct Lyndon subsequences

Theorem 6

$$TDS(\sigma,n) := \sum_{w\in\Sigma^n} (\text{ the number of distinct Lyndon subsequences in } w \ )$$
$$= \sum_{m=1}^{n} \left(L(\sigma,m) \sum_{k=m}^{n} \binom{n}{k} (\sigma-1)^{n-k} \right)$$

Expected distinct number $EDS(2,3)$ of Lyndon subsequences in a string is 20/8.

- $\Sigma = \{$a,b$\}$, $n = 3$

$\{$a,b$\}^3$	aaa	aab	aba	abb	baa	bab	bba	bbb	
distinct Lyndon subsequences	**a**aa	**a**ab aa**b** **a**a**b** **aab**	**a**ba a**b**a **ab**a	**a**bb a**b**b **a**b**b** **abb**	**b**aa b**a**a	**b**ab b**a**b b**ab**	**b**ba bb**a**	**b**bb	
the number	1	4	3	4	2	3	2	1	20

$TDS(2,3)$

Total number of distinct Lyndon subsequences

Theorem 6

$$TDS(\sigma, n) := \sum_{w \in \Sigma^n} (\text{ the number of distinct Lyndon subsequences in } w \text{ })$$

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

The above theorem means that the followings are equivalent:

A) Counting the number of distinct Lyndon subsequences in all strings w of length n.

B) Counting the number of strings w of length n that contain each Lyndon word as a subsequence.

Ideas for the proof of Theorem 6

Lyndon word of length ≤ 3

✓ ⋯ The string in Σ^n contains the Lyndon word as a substring.

Σ^3

	a	b	ab	aab	abb	# of ✓
aaa	✓	-	-	-	-	1
aab	✓	✓	✓	✓	-	4
aba	✓	✓	✓	-	-	3
abb	✓	✓	✓	-	✓	4
baa	✓	✓	-	-	-	2
bab	✓	✓	✓	-	-	3
bba	✓	✓	-	-	-	2
bbb	-	✓	-	-	-	1
# of ✓	7	7	4	1	1	20

A) the number of distinct Lyndon subsequences in all strings w of length n

$TDS(2,3)$

B) the number of strings w of length n that contain each Lyndon word as a subsequence

Ideas for the proof of Theorem 6

Lyndon word of length ≤ 3

✓ ⋯ The string in Σ^n contains the Lyndon word as a substring.

Σ^3

	a	b	ab	aab	abb	# of ✓
aaa	✓	-	-	-	-	1
aab	✓	✓	✓	✓	-	4
aba	✓	✓	✓	-	-	3
abb	✓	✓	✓	-	✓	4
baa	✓	✓	-	-	-	2
bab	✓	✓	✓	-	-	3
bba	✓	✓	-	-	-	2
bbb	-	✓	-	-	-	1
# of ✓	7	7	4	1	1	20

A) the number of distinct Lyndon subsequences in all strings w of length n

<u>Observation</u>

Two Lyndon words of equal length give the same value.

B) the number of strings w of length n that contain each Lyndon word as a subsequence

Proof of Theorem 6

- Let $Count(n, \Sigma, x)$ be the number of strings in Σ^n that contain $x \in \Sigma^m$ $(m \leq n)$ as a subsequence.

<u>Lemma 2</u>

For any $x_1, x_2 \in \Sigma^m$ and m, n $(m \leq n)$,

$$Count(n, \Sigma, x_1) = Count(n, \Sigma, x_2).$$

- $Count(3, \{\text{a, b}\}, \text{ba}) = 4$ {a**ba**, **ba**a, **ba**b, **b**b**a**}$\subseteq \{a,b\}^3$
- $Count(3, \{\text{a, b}\}, \text{aa}) = 4$ {**aa**a, b**aa**, **a**b**a**, **aa**b} $\subseteq \{a,b\}^3$

Induction

- $Count(l+1, \Sigma, yc)$ can be represented by $Count(l, \Sigma, y)$ and $Count(l, \Sigma, yc)$.

n \ m	1	⋯	k	k+1	⋯
1	✓	-	-	-	-
⋮	✓	✓	-	-	-
l	✓	✓	✓	✓	-
l+1	✓	✓	✓		
⋮	✓	✓	✓		

$k = |y| \ (y \in \Sigma^k, c \in \Sigma)$

# of strings in Σ^{l+1}
that contain $yc \in \Sigma^{k+1}$
as a subsequence

$$Count(l+1, \Sigma, yc) = \sigma\, Count(l, \Sigma, yc) + \left(Count(l, \Sigma, y) - Count(l, \Sigma, yc)\right)$$

of strings containing x

<u>Lemma 3</u>

Let $Count(n, \Sigma, x)$ be the number of strings in Σ^n that contain $x \in \Sigma^m$ $(m \leq n)$ as a subsequence. Then,

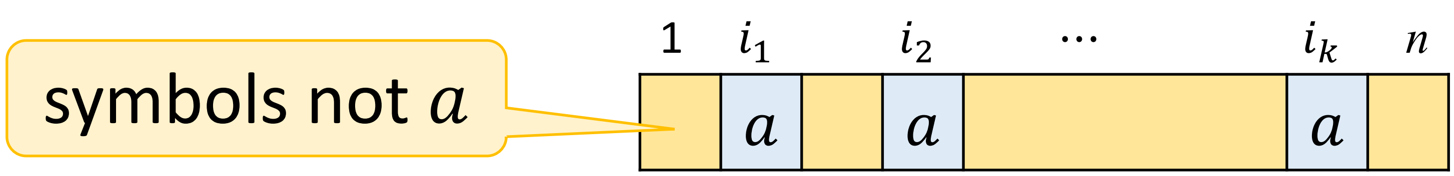
$$Count(n, \Sigma, x) = \sum_{k=m}^{n} \binom{n}{k} (\sigma - 1)^{n-k}.$$

- From Lemma 2,

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

the number of strings containing at least m a's

strings containing exactly k a's



Proof of Theorem 6

Lyndon word of length ≤ 3

Σ^3

	a	b	ab	aab	abb	# of ✓
aaa	✓	-	-	-	-	1
aab	✓	✓	✓	✓	-	4
aba	✓	✓	✓	-	-	3
abb	✓	✓	✓	-	✓	4
baa	✓	✓	-	-	-	2
bab	✓	✓	✓	-	-	3
bba	✓	✓	-	-	-	[illegible]
bbb	-	✓	-	-	[illegible]	1
# of ✓	7	7	4	1	1	20

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

$Count(3, \Sigma, ab) = 4$

Proof of Theorem 6

$\overbrace{\text{Lyndon word of length} \leq 3}$

Σ^3	a	b	ab	aab	abb	# of ✓
aaa	✓	-	-	-	-	1
aab	✓	✓	✓	✓	-	4
aba	✓	✓	✓	-	-	3
abb	✓	✓	✓	-	✓	4
baa	✓	✓	-	-	-	2
bab	✓	✓	✓	-	-	3
bba	✓	✓	-	-	-	[illegible]
bbb	-	✓	-	-	[illegible]	1
# of ✓	7	7	4	1	1	20

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

$$Count(3, \Sigma, ab) = 4$$

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

Exact values of $EDF(\sigma, n)$ and $EDS(\sigma, n)$

n \ σ	2	5
1	1.00	1.00
2	1.75	2.20
3	2.50	3.56
4	3.25	5.02
5	4.06	6.55
6	4.91	8.16
7	5.81	9.82
8	6.77	11.54
9	7.77	13.31
10	8.83	15.13

$EDF(\sigma, n)$

(expected # of distinct Lyndon **factors**)

n \ σ	2	5
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

$EDS(\sigma, n)$

(expected # of distinct Lyndon **subsequences**)

Conclusion

- We counted the number of Lyndon subsequences.

	factor	subsequence
max. distinct	$\binom{n+1}{2} - (\sigma - 1)\binom{q+1}{2} - r\binom{q+2}{2} + \sigma$	open
max. total	$\binom{n+1}{2} - (\sigma - 1)\binom{q+1}{2} - r\binom{q+2}{2} + n$	$2^n - (r+\sigma)2^q + n + \sigma - 1$
expected total	$\sum_{m=1}^{n} (L(\sigma, m)\,(n - m + 1)\,\sigma^{-m})$	$\sum_{m=1}^{n} \left(L(\sigma, m)\,\binom{n}{m}\,\sigma^{-m}\right)$
expected distinct	$\sum_{m=1}^{n} \left(L(\sigma, m) \sum_{k=1}^{\lfloor n/m \rfloor} (-1)^{k+1} \binom{n - km + k}{k} \sigma^{-km}\right)$	$\sum_{m=1}^{n} \left(L(\sigma, m) \sum_{k=m}^{n} \binom{n}{k}\,(\sigma - 1)^{n-k}\right) \sigma^{-n}$

Future work

- The value of $MDS(\sigma, n)$ is open.
- $MDS(\sigma, n) \coloneqq \max_{w \in \Sigma^n}($ # of distinct Lyndon subsequences in w $)$

n	$MDS(2,n)$	w
2	3	ab
3	4	abb
4	6	aabb
5	8	ababb
6	13	aababb
7	18	aaababb
8	28	aabababb

n	$MDS(2,n)$	w
9	41	aaabababb
10	63	aababababb
11	96	aaababababb
12	141	aabababababb
13	225	aaabababababb
14	335	aaababababbabb
15	538	aaababababababb

Each w is an instance of strings that achieve $MDS(\sigma, n)$ with $\sigma = 2$.

This sequence is not in OEIS.