Lyndon 分解と LZ 分解の関係性

～ Lyndon vs. Lempel-Ziv ～

中島祐人（九州大学）

joint work with

Juha Kärkkäinen, Dominik Kempa, Simon J. Puglisi
(University of Helsinki)

Arseny M. Shur
(Ural Federal University)

2016/11/21

文字列の分解

空でない部分文字列の列.

例えば...

a b a b a b a a b b b a a b a

a b,a b,a b,a a b b b,a a b,a

a b|a b|a b|a a b b b|a a b|a

特定の性質を満たす分解の例

- LZ 分解
- Lyndon 分解
- 回文分解

LZ77 分解 (自己参照なし)

定義 [Ziv and Lempel, 1977]

$t_1, t_2, \ldots, t_z$ が w の LZ77 分解 LZ_w であるとは，
次の条件をみたすことである．

- $w = t_1 t_2 \cdots t_k$;
- $t_1 = w[1]$;
- $t_1 \cdots t_{i-1} = w[1..j]$ とする，
  - $w[j+1]$ が $w[1..j]$ 中に出現しないとき，
    $t_i = w[j+1]$;
  - $w[j+1]$ が $w[1..j]$ 中に出現するとき，
    $t_i = w[j+1..j+q]$,
    $q = \max\{r \mid w[p..p+r-1] = w[j+1..j+r], p+r-1 \le j\}$.

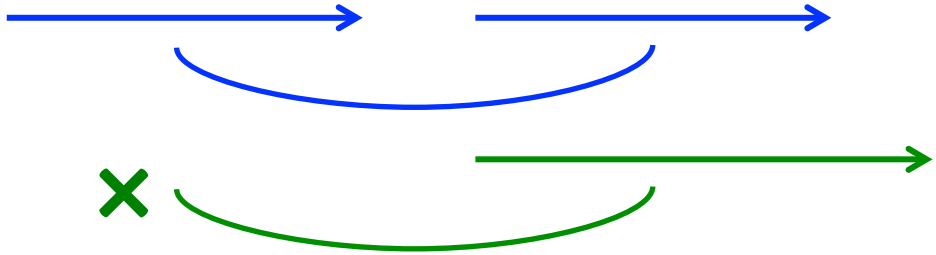
以降，単に LZ 分解とよぶ

LZ77 分解 (自己参照なし)

例

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

LZ_w = a | b | a b | a b a | a b | b | b a a b | a

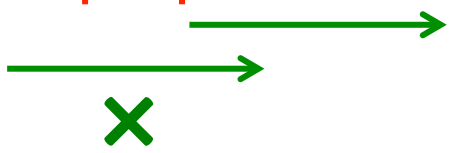


LZ77 分解 (自己参照なし)

例

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

LZ_w = a | b | a b a b a a b b b a a b a



参照先と factor 自身が重なることを認めない.

LZ77 分解 (自己参照なし)

例

$$w = \mathrm{a\ b\ a\ b\ a\ b\ a\ a\ b\ b\ b\ a\ a\ b\ a}$$

$$LZ_w = \mathrm{a|b|a\ b|a\ b\ a|a\ b|b|b\ a\ a\ b|a}$$

辞書式順序の定義

定義（辞書式順序）

任意の文字列 x, y について，x が y より辞書式順序が小さい（$x \prec y$）とは，次のいずれかの条件を満たすことである：
(1) $x[lcp(x, y)+1] \prec y[lcp(x, y)+1]$,
(2) x は y の真の接頭辞.

※ $lcp(x, y)$: x, y の最長共通接頭辞の長さ

例　(a ≺ b ≺ c)

a b b c ≺ a b c

a b ≺ a b c

Lyndon 文字列 [Lyndon, 1954]

定義

文字列 w が Lyndon 文字列であるとは，w のすべての真の接尾辞より w の辞書式順序が小さいことである．

※ 真の接尾辞 : もとの文字列自身でない接尾辞．

w							w の真の接尾辞
a	a	b	a	b	b	$<$	a b a b b
						$<$	b a b b
						$<$	b a b b
						$<$	b a b b
						$<$	b a b b

Lyndon 文字列

Lyndon 分解 [Chen, Fox, Lyndon, 1958]

定義

文字列の列 $u_1^{p_1},\ldots,u_m^{p_m}$ が w の Lyndon 分解 LF_w であるとは，次の条件を満たすことである：
$u_1 > \ldots > u_m$ が Lyndon 文字列，$p_i \geq 1\ (1 \leq i \leq m)$，
かつ $w = u_1^{p_1}\ldots u_m^{p_m}$．

$$w = \underset{u_1}{\mathrm{a\,b\,c}}\,|\,\underset{u_2}{\mathrm{a\,b\,b}}\,|\,\underset{u_2}{\mathrm{a\,b\,b}}\,|\,\underset{u_3}{\mathrm{a\,a\,b\,c}}\,|\,\underset{u_4}{\mathrm{a}}\,|\,\underset{u_4}{\mathrm{a}}\,|\,\underset{u_4}{\mathrm{a}}$$

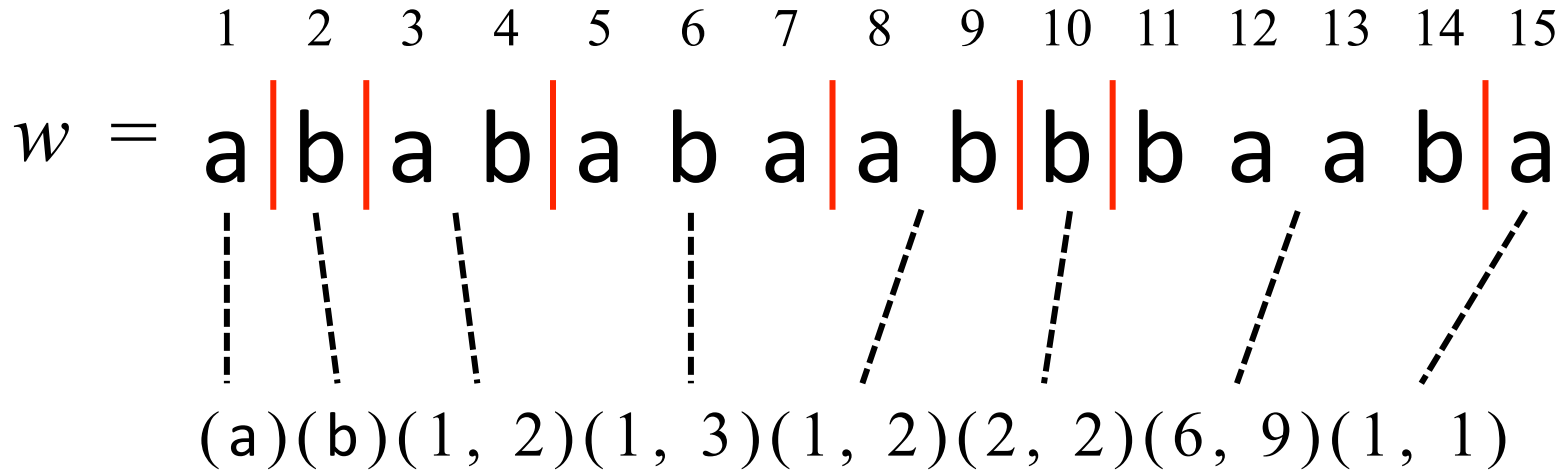
$$LF_w = \underset{u_1^1}{(\mathrm{abc})^1}\,|\,\underset{u_2^2}{(\mathrm{abb})^2}\,|\,\underset{u_3^1}{(\mathrm{aabc})^1}\,|\,\underset{u_4^3}{(\mathrm{a})^3}$$

$|LF_w|$: Lyndon 分解のサイズ(= factor の数)

分解を考えること

- 文字列の構造を捉える
- 文字列の圧縮表現

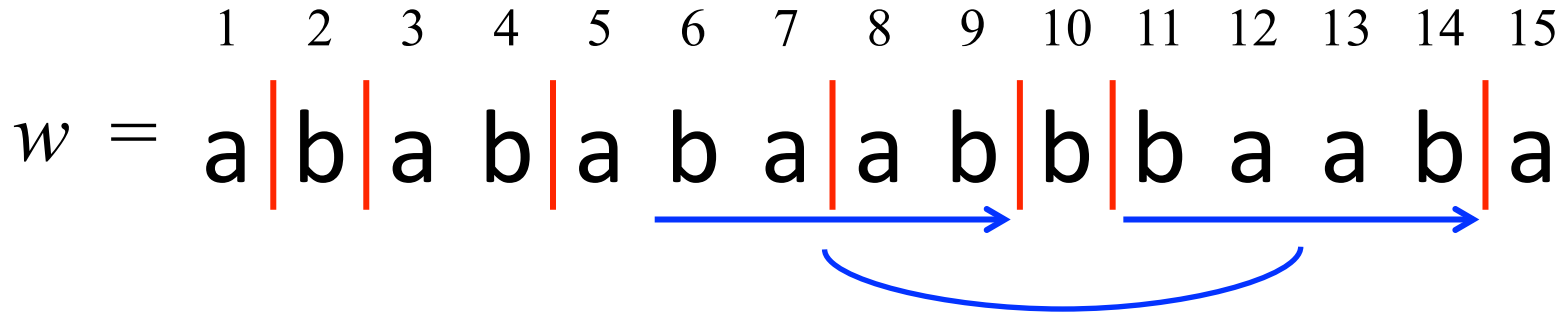
  例 LZ 分解



分解を考えること

- 文字列の構造を捉える
- 文字列の圧縮表現

  例 LZ 分解



(a)(b)(1, 2)(1, 3)(1, 2)(2, 2)(6, 9)(1, 1)

分解を考えること

- 文字列の構造を捉える
- 文字列の圧縮表現

  例 LZ 分解

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

$w =$ a | b | a b | a b a | a b | b | b a a b | a

(a)(b)(1, 2)(1, 3)(1, 2)(2, 2)(6, 9)(1, 1)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

$w =$ a b a b a b a a b b b a a b a

分解を考えること

- 文字列の構造を捉える
- 文字列の圧縮表現

  例 LZ 分解

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

$w =$ a | b | a b | a b a | a b | b | b a a b | a

(a)(b)(1, 2)(1, 3)(1, 2)(2, 2)(6, 9)(1, 1)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

$w =$ a b a b a b a a b b b a a b a

最小文法の下界

Lyndon 分解と LZ 分解は,
最小文法の下界であることが知られている.

LZ 分解
The Smallest Grammar Problem

Moses Charikar, Eric Lehman, April Lehman, Ding Liu, Rina Panigrahy, Manoj Prabhakaran, Amit Sahai, abhi shelat, 2005

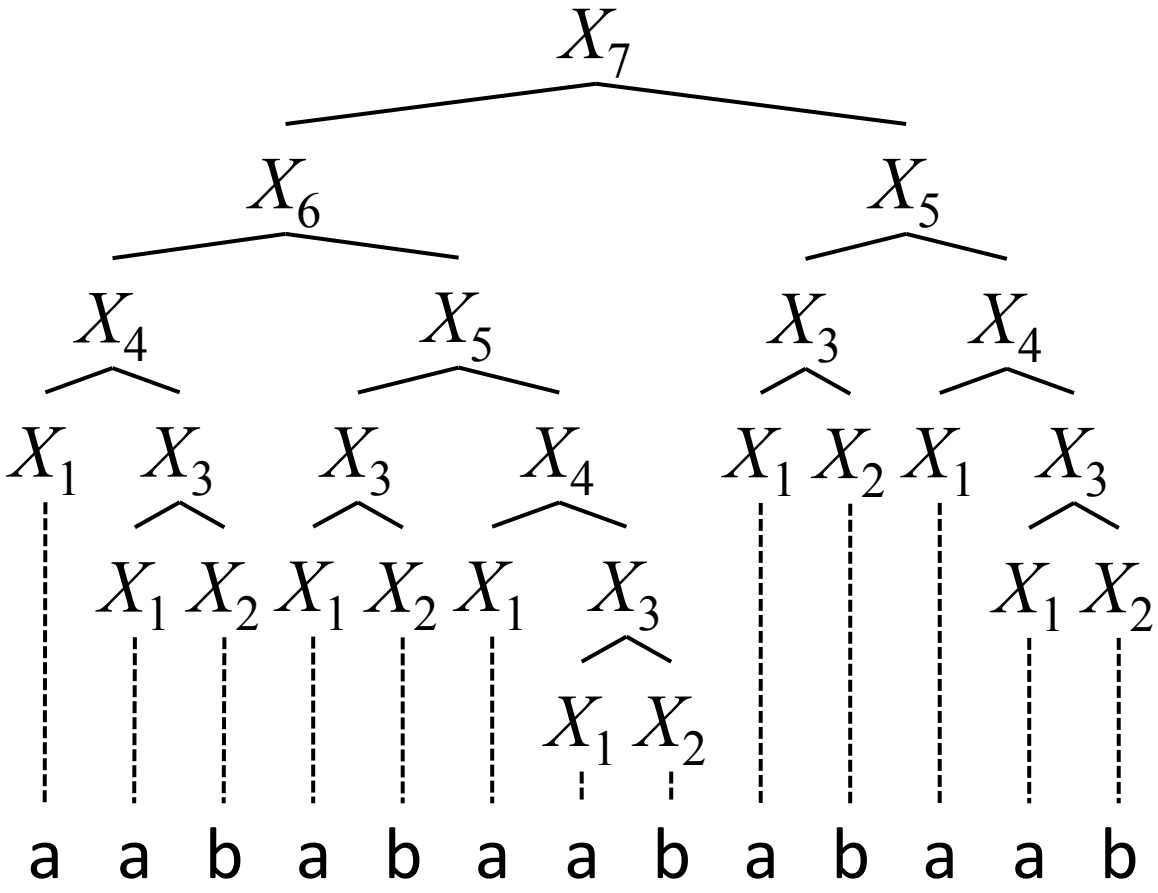
Lyndon 分解
Faster Lyndon Factorization Algorithms for SLP and LZ78 Compressed Text

Tomohiro I, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, TCS 2016, SPIRE 2013

最小文法問題

単一の文字列を導出する最小の文脈自由文法を求める問題.

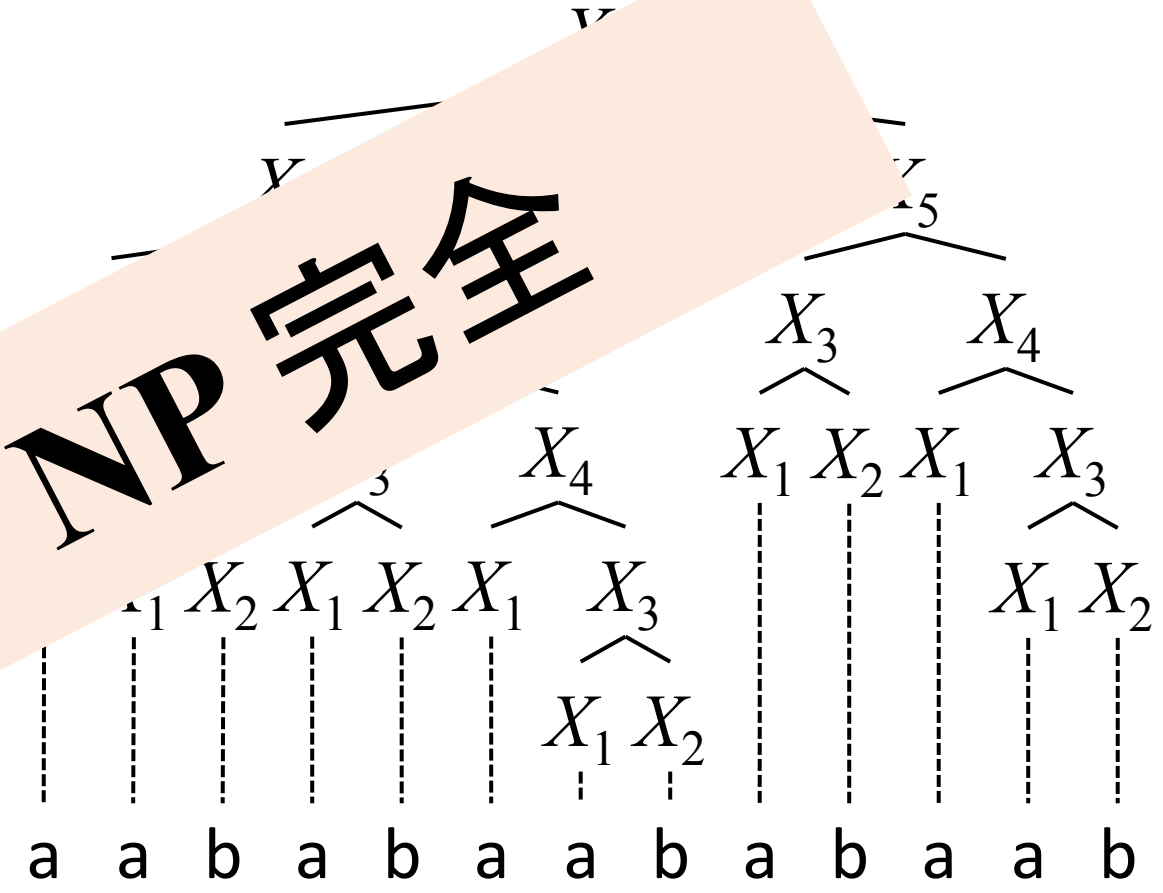
$X_1 \to \mathrm{a}$
$X_2 \to \mathrm{b}$
$X_3 \to X_1\, X_2$
$X_4 \to X_1\, X_3$
$X_5 \to X_3\, X_4$
$X_6 \to X_4\, X_5$
$X_7 \to X_6\, X_5$



最小文法問題

単一の文字列を導出する最小の文脈自由文法を求める問題.

$X_1 \to a$
$X_2 \to b$
$X_3 \to X_1 X_2$
$X_4 \to X_$
$X_5 \to X_3$
$X_6 \to X_4 \lambda$
$X_7 \to X_6 X_5$

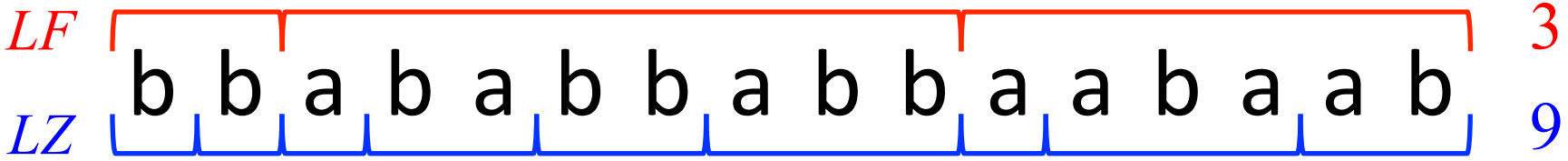


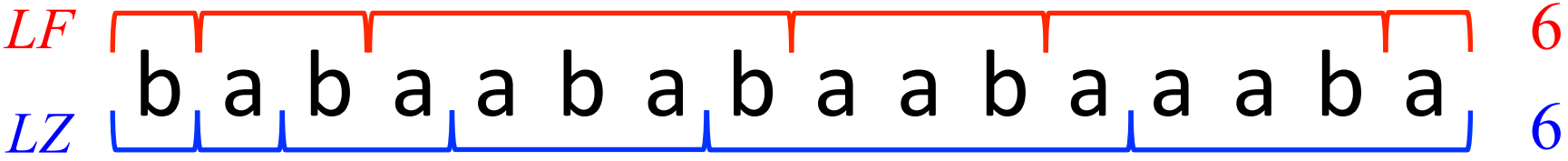
NP 完全

Lyndon vs. LZ 予想

Lyndon 分解のサイズと LZ 分解のサイズに関係は
あるのだろうか？

予想 [Nakashima, 2014]
任意の文字列 w について, $\lvert LF_w \rvert \leq \lvert LZ_w \rvert$.

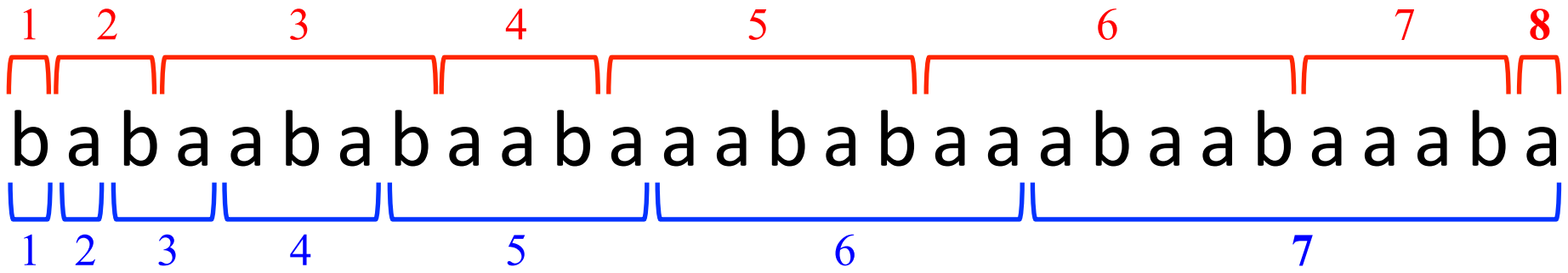




Lyndon vs. LZ 予想

予想 [Nakashima, 20

任意の文字列 w につ $|LF_w$.



- アルファベットサイズ 2, 長さ 30 以下
- アルファベットサイズ 3, 長さ 19 以下
- アルファベットサイズ 4, 長さ 15 以下

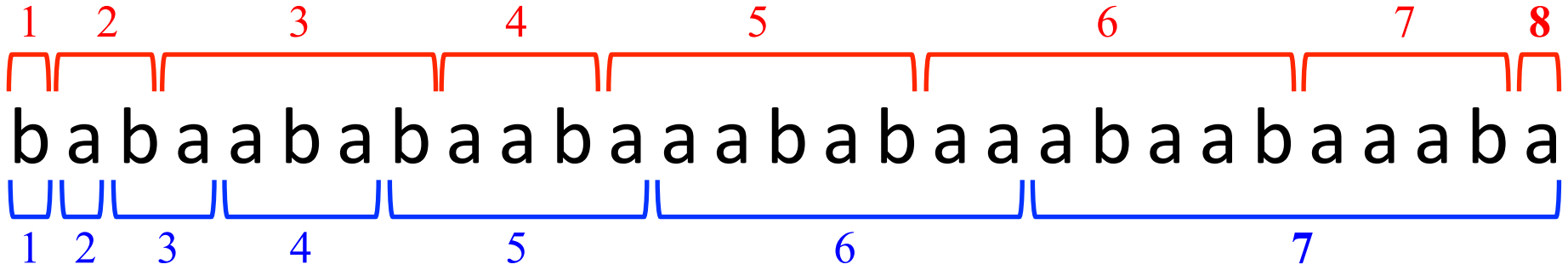
以上の文字列の中で唯一の反例.

反例文字列の特徴

反例文字列は以下のように表すことができ，
$|LF_w| - |LZ_w| \to \infty$ となる文字列を生成できる．

$$B_k = (\mathrm{a}^k\mathrm{ba}^1\mathrm{b})\,(\mathrm{a}^k\mathrm{ba}^2\mathrm{b})\,\ldots\,(\mathrm{a}^k\mathrm{ba}^{k-1}\mathrm{b})\,(\mathrm{a}^k\mathrm{b})$$

$$X_k = B_0B_1\cdots B_k\,\mathrm{a}$$



（反例は文字列 X_3）

$$|LF_{X_k}| - |LZ_{X_k}| = k - 2\ (k \geqq 3).$$

主結果

定理(Upper bound)
任意の文字列 w について, $\lvert LF_w \rvert < 2\lvert LZ_w \rvert$.

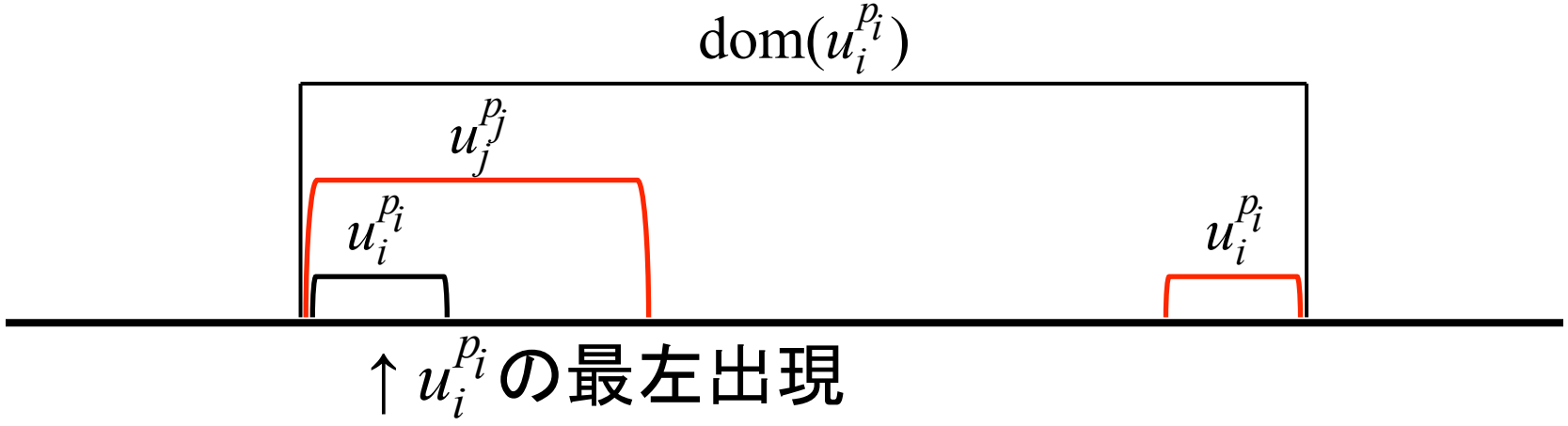
Key lemma

補題

任意の Lyndon factor $u_i^{p_i}$ の domain $\mathrm{dom}(u_i^{p_i})$ について，$\mathrm{dom}(u_i^{p_i})$ は少なくとも「$|\mathrm{dom}(u_i^{p_i})| / 2\urcorner + 1$ 個の LZ factor の開始位置を含む.

※ $\mathrm{dom}(u_i^{p_i}) = u_j^{p_j} \ldots u_i^{p_i}$ (s.t. $u_i^{p_i}$ の最左出現が u_j の接頭辞).

※ $|\mathrm{dom}(u_i^{p_i})| = i - j$.



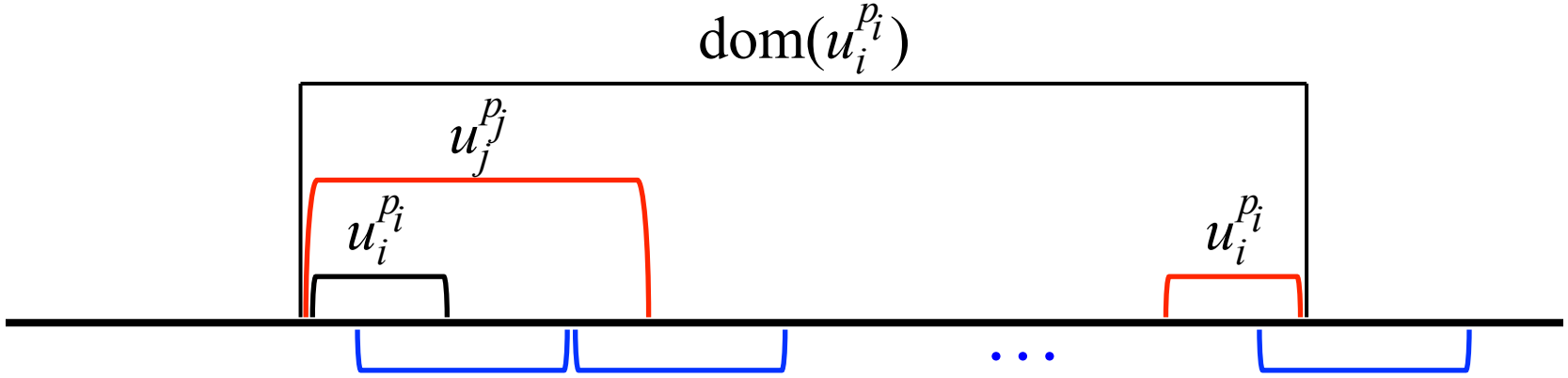
Key lemma

補題

任意の Lyndon factor $u_i^{p_i}$ の domain $\mathrm{dom}(u_i^{p_i})$ について，$\mathrm{dom}(u_i^{p_i})$ は少なくとも「$|\mathrm{dom}(u_i^{p_i})| / 2\rceil + 1$ 個の LZ factor の開始位置を含む.

※ $\mathrm{dom}(u_i^{p_i}) = u_j^{p_j} \ldots u_i^{p_i}$ (s.t. $u_i^{p_i}$ の最左出現が u_j の接頭辞).

※ $|\mathrm{dom}(u_i^{p_i})| = i - j$.



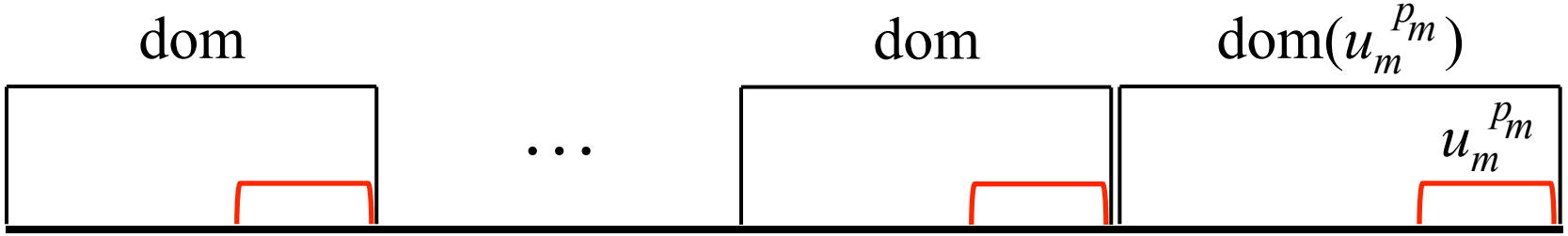
少なくとも「$(i-j) / 2\rceil + 1$ 個の LZ factor

Key lemma

- Lyndon 分解を domain の列に分けて補題を適用.

補題

任意の Lyndon factor $u_i^{p_i}$ の domain $\mathrm{dom}(u_i^{p_i})$ について, $\mathrm{dom}(u_i^{p_i})$ は少なくとも「$|\mathrm{dom}(u_i^{p_i})| / 2\rceil + 1$ 個の LZ factor の開始位置を含む.



定理 (Upper bound)

任意の文字列 w について, $|LF_w| < 2|LZ_w|$.

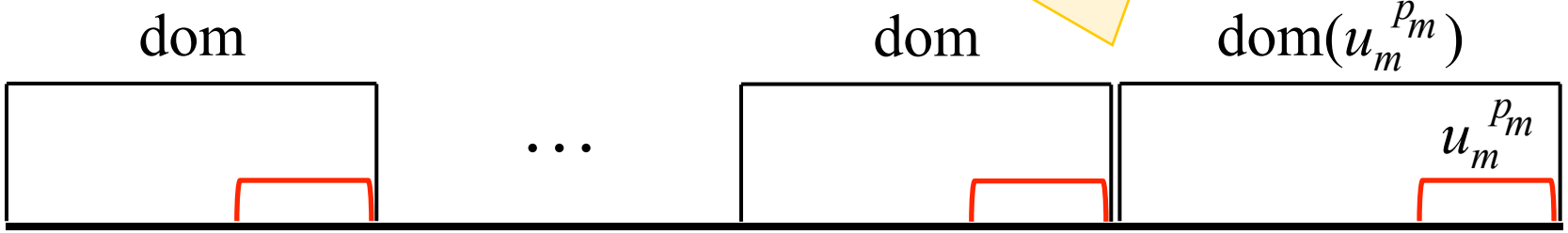
Key lemma

- Lyndon 分解を domain の列に分けて補題を適用.

補題

任意の Lyndon factor $u_i^{p_i}$ の d
$\mathrm{dom}(u_i^{p_i})$ は少なくとも「$|\mathrm{dom}(u_i^{p_i})|$
開始位置を含む.

$$\lceil x/2 \rceil + \lceil y/2 \rceil + 2 \geq \lceil (x+y+1)/2 \rceil + 1$$



定理 (Upper bound)

任意の文字列 w について, $|LF_w| < 2|LZ_w|$.

まとめと今後の課題

定理 (Upper bound)
任意の文字列 w について，$\lvert LF_w\rvert < 2\lvert LZ_w\rvert$.

定理 (Lower bound)
$\lvert LF_w\rvert = \lvert LZ_w\rvert + \Theta(\sqrt{\lvert LZ_w\rvert})$ となる文字列 w が存在する. ※ $\lvert LF_w\rvert \leq \lvert LZ_w\rvert$ 予想の反例文字列.

今後の課題

- よりタイトな上界と下界の証明.
- 自己参照を許した LZ 分解での予想，証明.