ERATO湊離散処理構造系プロジェクト
2013年度初夏のワークショップ

圧縮文字列上のLyndon分解

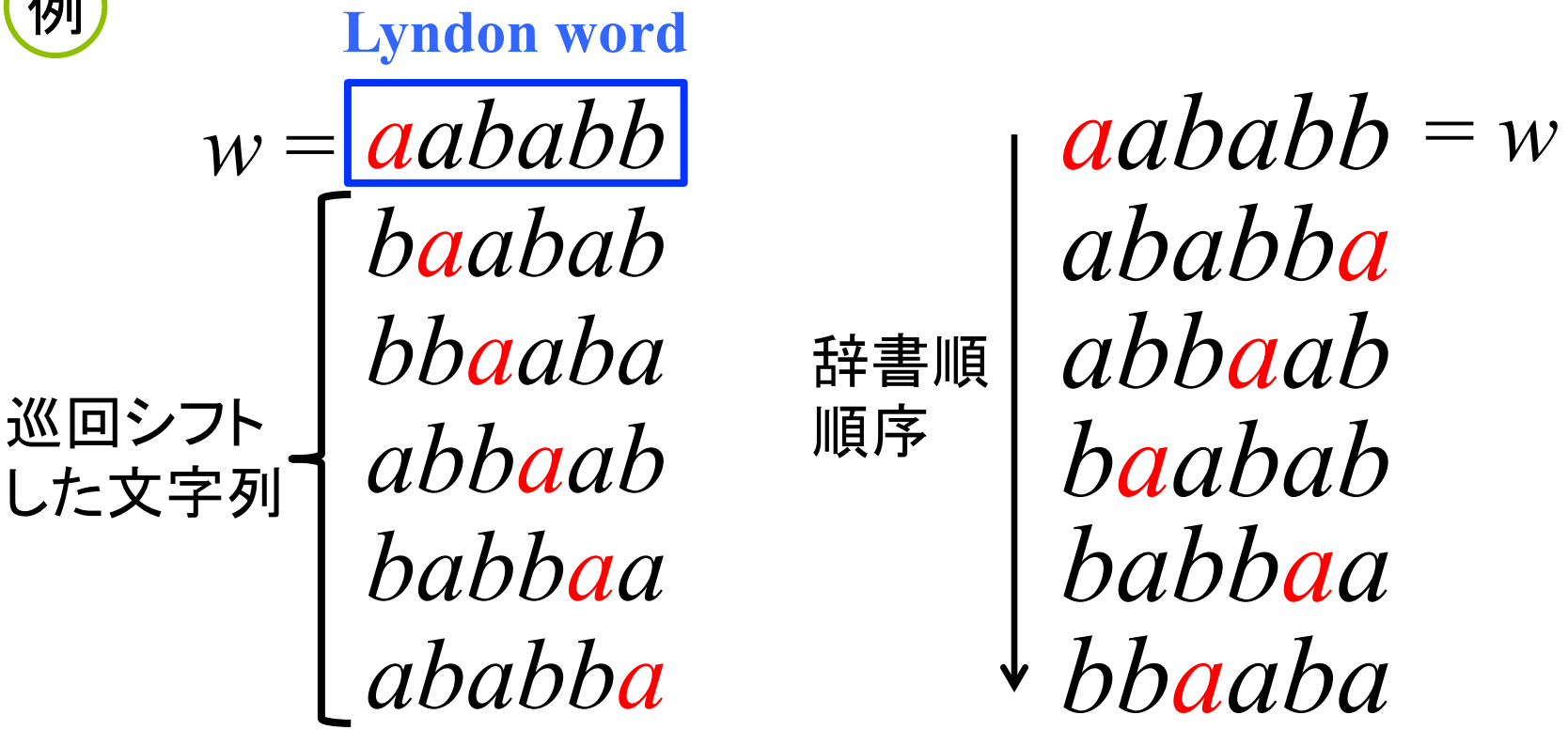
九州大学
井　智弘，中島祐人，稲永俊介，坂内英夫，竹田正幸

Lyndon word

[R. C. Lyndon, 1954]

文字列 w が Lyndon word であるとは, w の辞書式順序が, w を巡回シフトしたいずれの他の文字列よりも真に小さいこと.

例



Lyndon 分解

[K. T. Chen, R. H. Fox and R. C. Lyndon., 1958]

定義

$LF_w = l_1^{p_1} \ldots l_m^{p_m}$ が文字列 w の Lyndon 分解であるとは，
次の条件を満たすことである．

$p_i \geq 1\ (1 \leq i \leq m)$, $l_1 > l_2 > \ldots > l_m$ は Lyndon word.

例

$$w = a\,b\,c\,|\,a\,b\,b\,|\,a\,b\,b\,|\,a\,a\,b\,c\,|\,a\,|\,a\,|\,a$$

$$LF_w = (abc)(abb)^2(aabc)(a)^3$$

factor とよぶ．

いずれも Lyndon word

factor の辞書式順序　$abc > abb > aabc > a$

任意の文字列に対して，その Lyndon 分解は一意である．

Lyndon word, Lyndon 分解の応用

- 以下のように様々な応用がされている.

Lyndon word

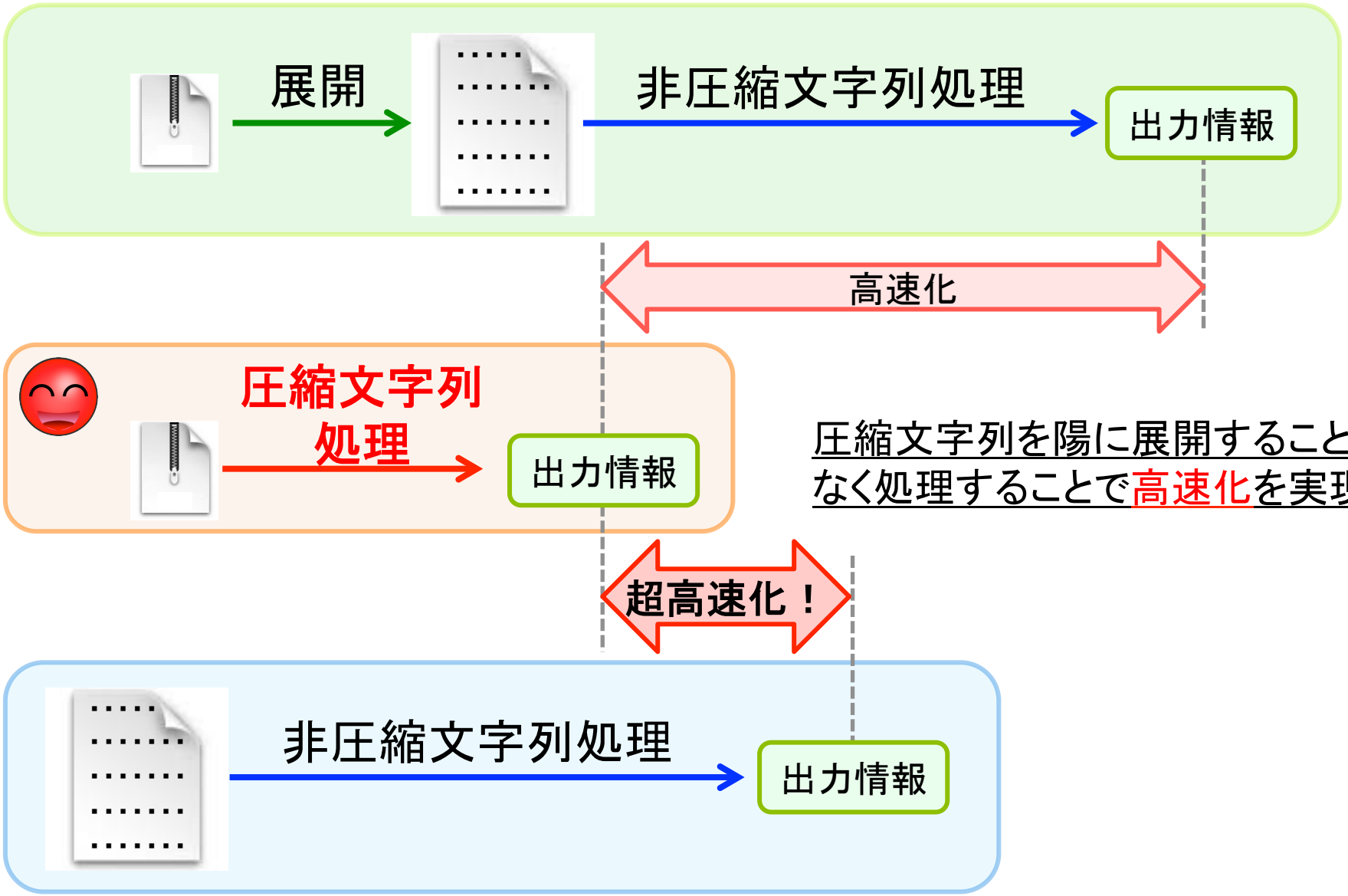
- 文字列照合
- 文字列組み合わせ論
- リー代数
- etc.

Lyndon 分解

- bijective Burrows-Wheeler transform [M. Kufleitner, 2009]
- 離散幾何学
- etc.

Lyndon word, Lyndon 分解の研究は重要である.

研究背景



圧縮文字列を陽に展開することなく処理することで高速化を実現.

文字列の圧縮表現

Straight-line Program (SLP)

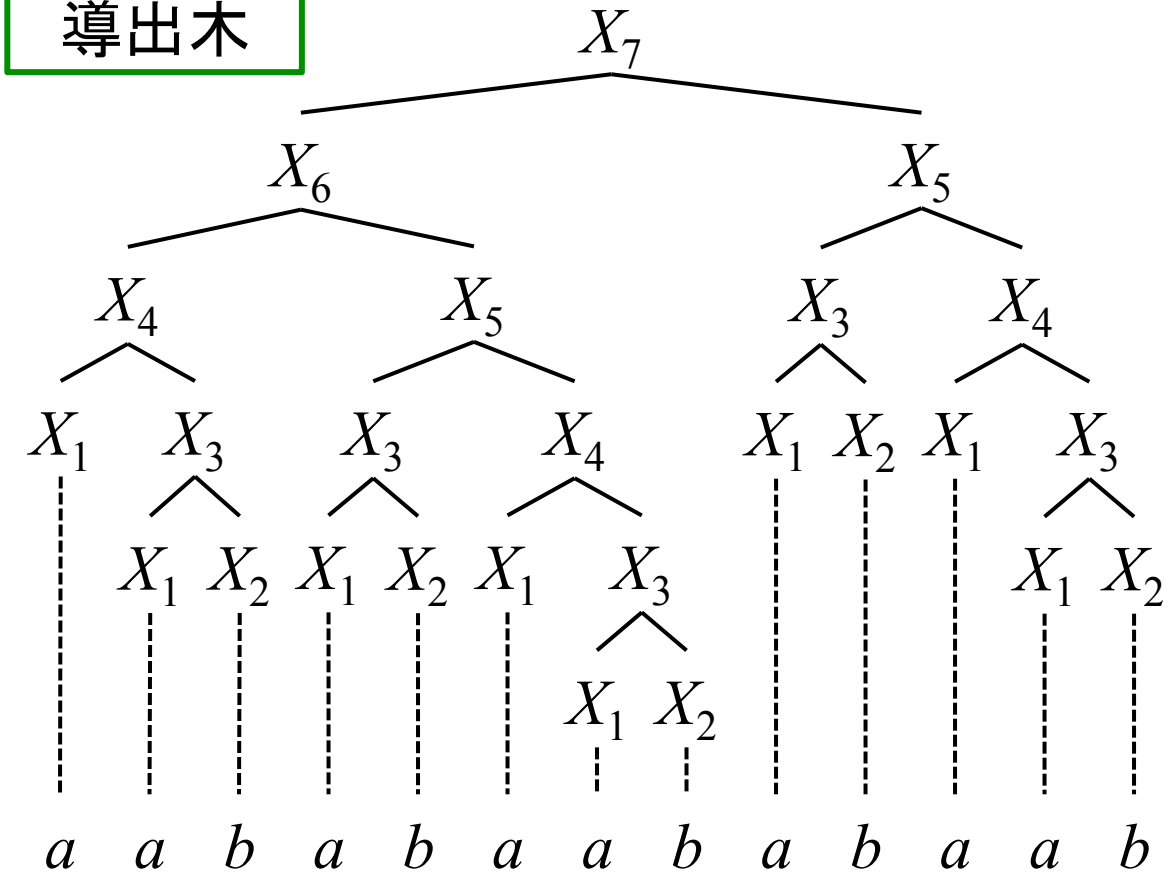
- 単一の文字列を導出するチョムスキー標準形の文脈自由文法.

例

SLP

$$
\begin{aligned}
X_1 &\to a \\
X_2 &\to b \\
X_3 &\to X_1\, X_2 \\
X_4 &\to X_1\, X_3 \\
X_5 &\to X_3\, X_4 \\
X_6 &\to X_4\, X_5 \\
X_7 &\to X_6\, X_5
\end{aligned}
$$

導出木



本研究における問題と目標

問題

文字列 w を表現する SLP S があたえられたとき，
w に対する Lyndon 分解 $LF(w)$ を計算する．

目標

n に対する多項式時間で $LF(w)$ を計算する．
（n は SLP のサイズ）

- 長さ N の文字列に対する $LF(w)$ を $O(N)$ 時間で計算できることが知られている．[J. P. Duval, 1983]
- SLP の展開長 $N = O(2^n)$ であるため，高い圧縮率の SLP においては，n の多項式時間で計算することができれば，非圧縮文字列に対する計算より高速である．

研究成果 1

CPM 2013 に採択（先週発表）

文字列 w を導出するサイズ n の SLP S があたえられたとき，$LF(w)$ を $O(mn^4)$ 時間，$O(n^2)$ 領域で計算できる．
ただし，m は Lyndon factor の数．

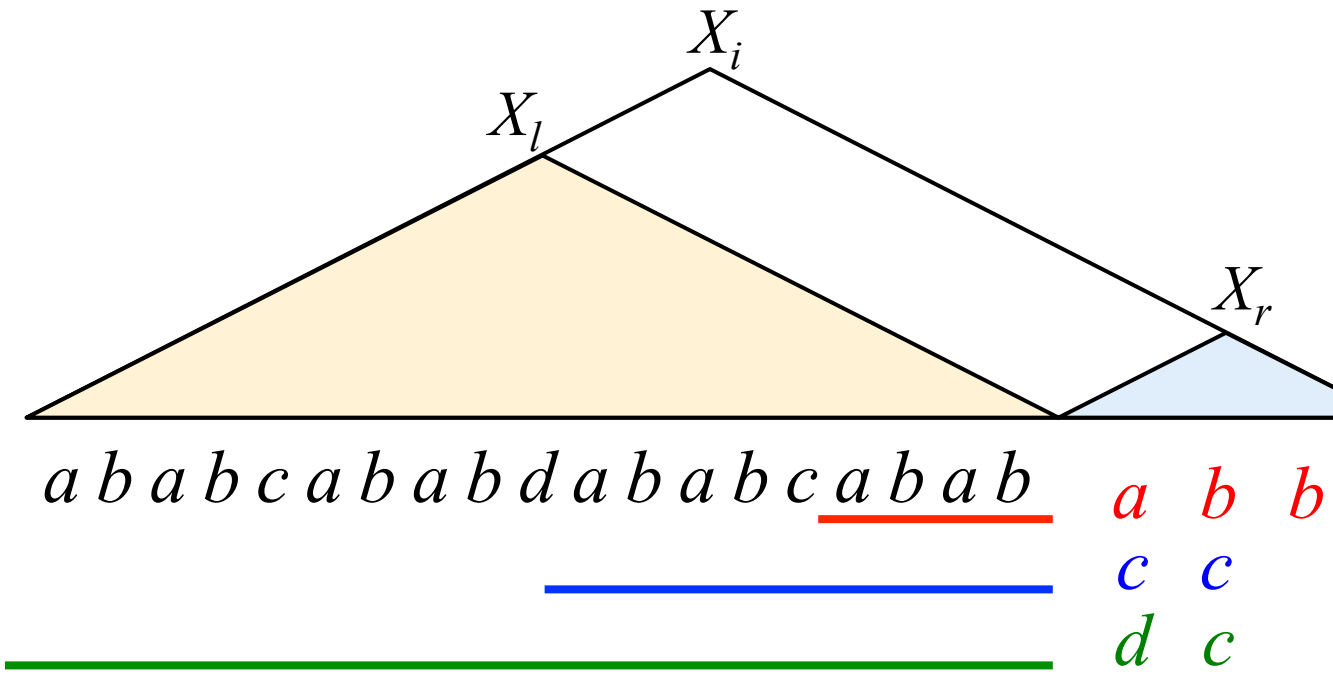
- $LF(w)$ の最後の factor は，w の辞書式順序最小の suffix である．[J. P. Duval, 1983]
- この性質から，以下の部分問題を考えることにより，SLP で表現された w に対して $LF(w)$ を計算できる．

部分問題

文字列 w を導出する SLP S があたえられたとき，w の辞書式順序最小の suffix を計算する．

主要な考え方

- 部分問題を解くために，後ろに文字列を連結したときに辞書式順序最小となる suffix の集合を考える.



- SLP の変数は2つの変数の連結であるため，上のような集合が計算された2つの変数から連結した変数に対する集合を計算する.

研究成果 2

SPIRE 2013 に採択（一昨日）

文字列 w を導出するサイズ n , 導出木の高さ h の SLP S があたえられたとき, $LF(w)$ を $O(nh(n + \log N \log n))$ 時間, $O(n^2)$ 領域で計算できる.

- 2つの文字列 u, v に対して, $LF(u), LF(v)$ が計算されているとき, $LF(uv)$ を計算する.
- $LF(uv)$ の切れ目は, $LF(u), LF(v)$ の factor を分割することはない. [J. W. Daykin, et al., 1983][A. Apostolico, et al., 1995]

主要な考え方

- SLP の変数 $X_i \to X_l X_r$ について, X_l, X_r が導出する文字列に対する Lyndon 分解 $LF(X_l), LF(X_r)$ があたえられたとき, $LF(X_i)$ を計算する.