

夏の LA シンポジウム 2026

部分文字列方程式による文字列表現

柴田 紘希（九州大学）

坂内 英夫（東京科学大学）



発表スライド

- 文字列を省領域で表現する手法（**圧縮手法**）はたくさんある
- 多くの圧縮手法は**部分文字列の一致関係**を用いている

- 文字列を省領域で表現する手法（**圧縮手法**）はたくさんある
- 多くの圧縮手法は**部分文字列の一致関係**を用いている

research question

部分文字列の一致関係をベースとする中で最も「自然な」圧縮手法は何か？

- 文字列を省領域で表現する手法（**圧縮手法**）はたくさんある
- 多くの圧縮手法は**部分文字列の一致関係**を用いている

research question

部分文字列の一致関係をベースとする中で最も「自然な」圧縮手法は何か？

研究で行ったこと

部分文字列方程式を用いた文字列圧縮手法を新たに考え、その性能を評価した

目次

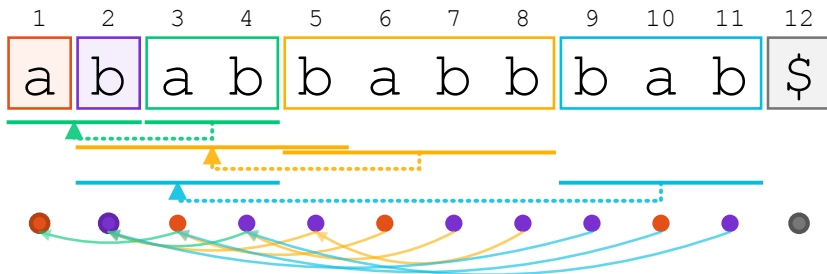
1. 部分文字列方程式 (SES)
2. Suffixient set サイズの SES の構成
3. SES・BMS のサイズ関係

圧縮手法の代表例: LZ 分解

2/22

LZ 分解: 文字列を以下のフレーズ列に分解して表す圧縮手法

1. **char phrase:** 単一の文字を表す
2. **copy phrase:** より左側に現れる同じ部分文字列の参照を表す

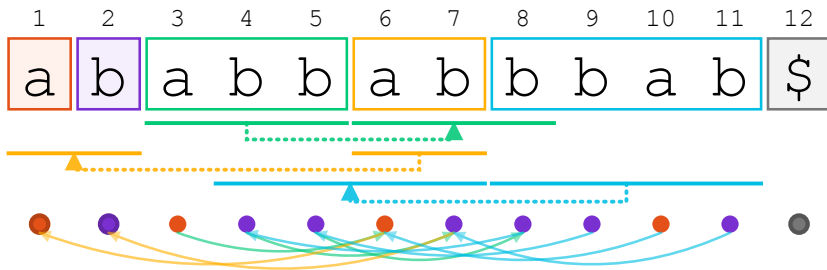


LZ分解の一般化: 双方向スキーム

3/22

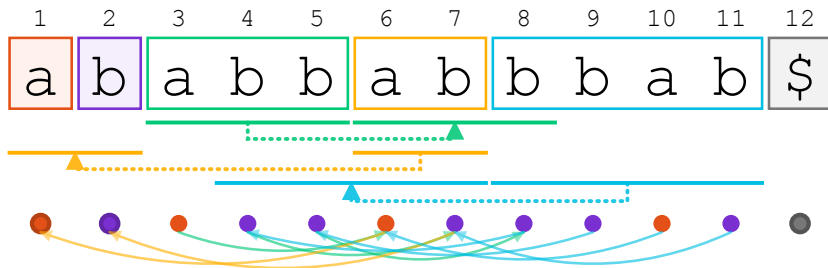
Bidirectional Macro Scheme (BMS): copy phraseが自身より右側の文字列も参照できるようなLZ分解の一般化

- ループができてしまうような参照関係は許さない



BMSは**部分文字列の一致関係**を活かすことで文字列を圧縮しているが、以下のような制約が存在する

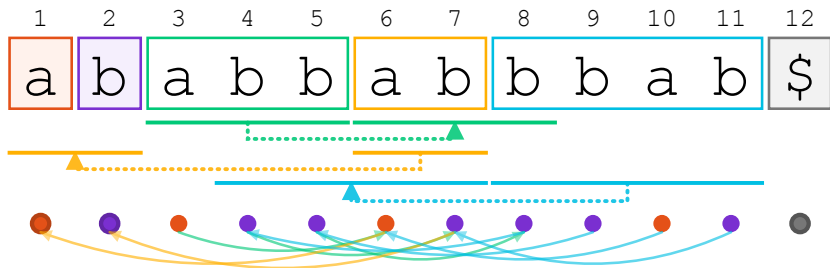
- 参照関係が向きづけを持つ
- 1つの text position が複数の text position を参照できない



BMSは**部分文字列の一致関係**を活かすことで文字列を圧縮しているが、以下のような制約が存在する

- 参照関係が向きづけを持つ
- 1つの text position が複数の text position を参照できない

→ **これらの制約を必要としない**文字列表現手法を考えられるのでは？



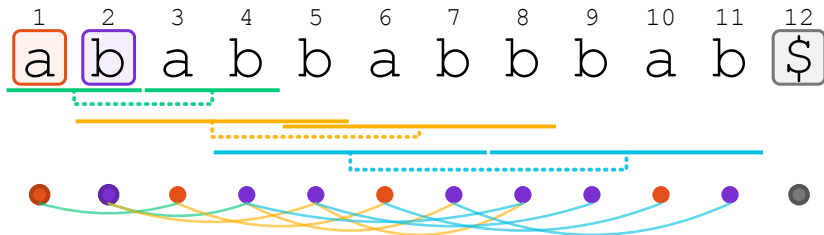
部分文字列方程式 (SE)

5/22

部分文字列方程式 (Substring Equation System, SES): 文字列を以下の形の制約の集合によって記述する表現手法

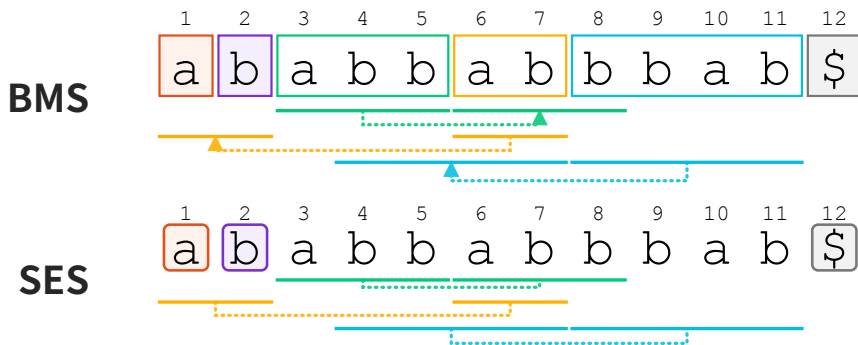
1. Equality 制約 $(i, j, \ell): w[i..i + \ell - 1] = w[j..j + \ell - 1]$
2. Assignment 制約 $(k, c): w[k] = c$

w を表す **SES**: 全制約を満たす文字列が w のみである SES



観察 1: BMS と SES の関係

w に対する phrase 数 t の BMS が与えられたとき、 w を表す制約数 t の SES を構成できる



- 逆問題（与えられた条件を満たす文字列を構成する問題）の文脈でよく用いられている
 - 例: ボーダー配列・極大繰り返しの集合などからの文字列復元
[Gawrychowski et al., 2020]
- 文字列長＋制約数に対して線形時間で解を求められ、解の一意性も判定できる [Gawrychowski et al., 2020]
- 文字列の圧縮表現手法としての研究は存在していなかった

部分文字方程式を考える利点

8/22

直感: SES は BMS よりシンプルな構造なため、扱いやすそう

- フレーズの重なりを許す
- 参照方向の向きづけが存在しない

直感: SES は BMS よりシンプルな構造なため、扱いやすそう

- フレーズの重なりを許す
- 参照方向の向きづけが存在しない

実際に得られた成果: SES を用いることで、最小 suffixient set サイズ $\chi(w)$ に対して $O(\chi(w))$ words で文字列を表現する手法を与えた

- $O(\chi(w))$ サイズの文字列表現ができるかは未解決だった
- BMS では扱いづらいところを SES の形で扱うことで解決

目次

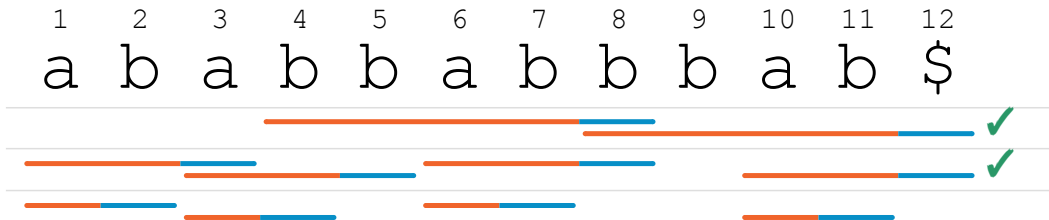
1. 部分文字列方程式 (SES)
2. Suffixient set サイズの SES の構成
3. SES・BMS のサイズ関係

right extension

9/22

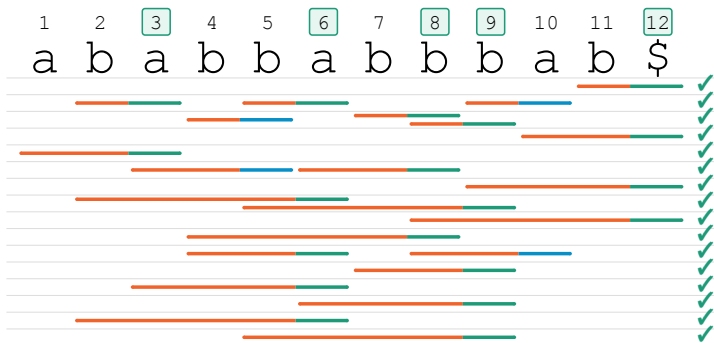
定義 1: right extension (RE)

文字列 w の部分文字列 x について、 x の直後に現れる文字 c が2通り以上存在するとき、各 xc を w の **right extension (RE)** と呼ぶ。



定義 2: suffixient set [Depuydt et al., 2024]

任意の RE について、その出現終了位置の少なくとも一カ所が S に含まれるよう text position の集合 S を w の **suffixient set** と呼ぶ。



定義 2: suffixient set [Depuydt et al., 2024]

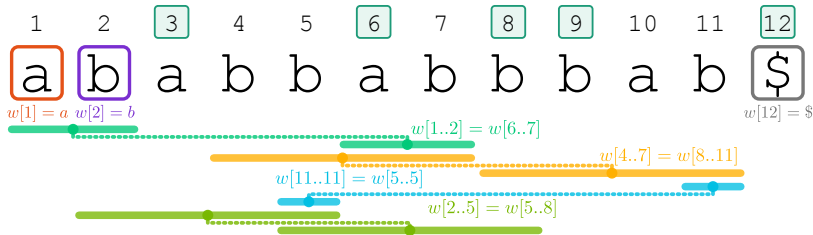
任意の RE について、その出現終了位置の少なくとも一カ所が S に含まれるような text position の集合 S を w の **suffixient set** と呼ぶ。

- 最小 suffixient set を 1 つ固定して S とおき、そのサイズを $\chi(w) = |S|$ と表す
- $\chi(w)$ の値は、文字列によっては非常に小さくなる
- 任意の文字列 w を $O(\chi(w))$ words で表現できるかは未知だった

定理 1: $O(\chi(w))$ サイズの部分文字列方程式

任意の文字列 w に対して、 w を表す制約数 $\chi(w) + \sigma(w) - 1 \leq 2\chi(w)$ の SES を構成できる^a。

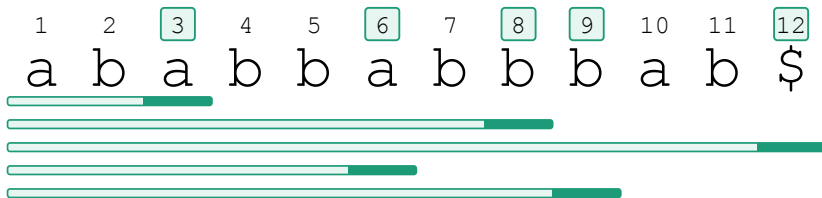
^a $\sigma(w)$: w 中の相異なる文字の個数



suffixient set への順序付け

13/22

S の各要素 $i \in S$ を、 $i - 1$ 文字目までの接頭辞を反転した文字列 $x[1..i - 1]^R$ の辞書順に並べ、この列を $s_1, \dots, s_{\chi(w)}$ とおく

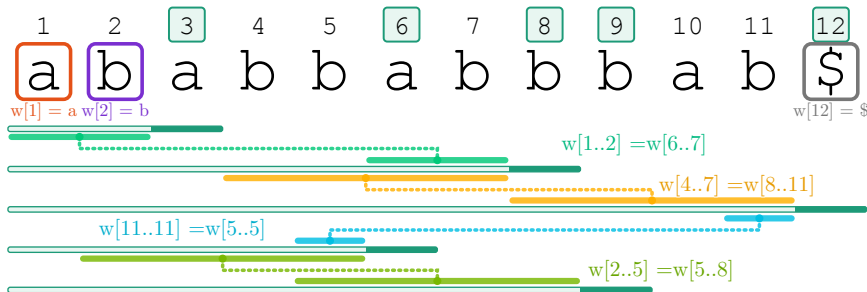


部分文字列方程式の構成

14/22

部分文字列方程式 $\Lambda(w)$ の構成:

1. w 中の各文字に1つずつ、合計 $\sigma(w)$ 個の Assignment 制約
2. s_i, s_{i+1} について $w[1..s_i - 1]$ と $w[1..s_{i+1} - 1]$ の共通接尾辞を表す
合計 $\chi(w) - 1$ 個の Equality 制約

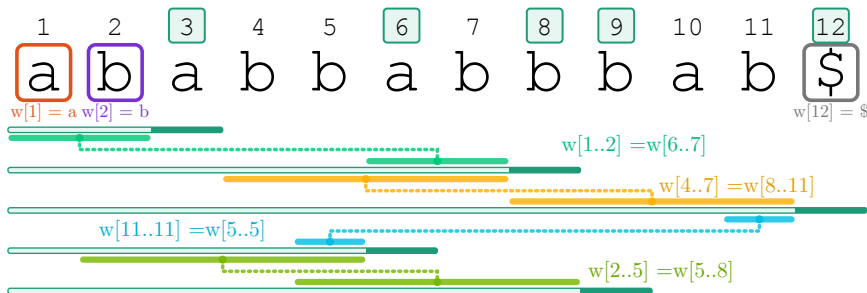


共通接尾辞部分の一致関係

15/22

観察: $\Lambda(w)$ の Equality 制約のみから、任意の $s_i, s_j \in S$ について $w[1..s_i - 1], w[1..s_j - 1]$ の共通接尾辞の一致関係を導ける

- 隣接部分の一致関係+推移律 より成り立つ

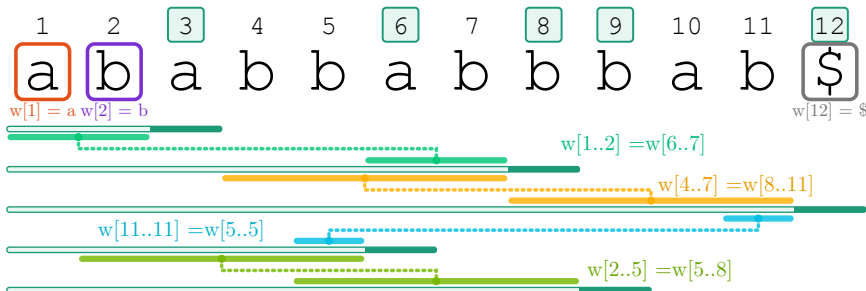


正当性の証明 (1)

16/22

正当性の証明: 任意の同一部分文字列間の一致関係が今回の Equality 制約のみから導けることを示す

証明の方針: 一致する部分文字列の長さ ℓ の降順に帰納法



正当性の証明 (2)

17/22

目標: 長さ ℓ の文字列 u の出現の組 $w[i - \ell..i - 1] = w[j - \ell..j - 1] = u$ の一致関係を Copy 制約と帰納法の仮定のみから導く

u の各出現の直後の文字 $c = w[i], c' = w[j]$ について場合分け:

(1) $c = c'$ の場合: $uc = uc'$ は長さ $\ell + 1$ の一致関係なので、仮定よりこれらの一致関係は $\Delta(w)$ から導ける

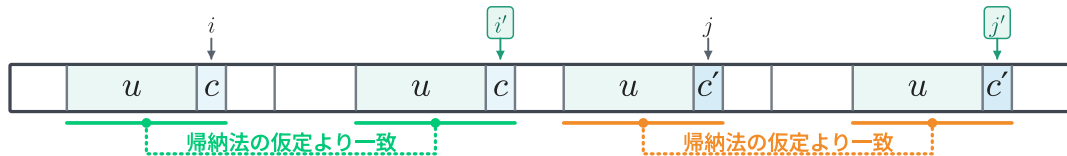


正当性の証明 (3)

18/22

(2) $c \neq c'$ の場合:

1. $c \neq c'$ より uc, uc' は RE
2. suffixient set の定義より、suffixient set の位置で終わるような uc, uc' の出現が必ず存在 (下図の i', j')
3. 帰納法の仮定より、これらの uc, uc' の一致関係は $\Lambda(w)$ から導ける

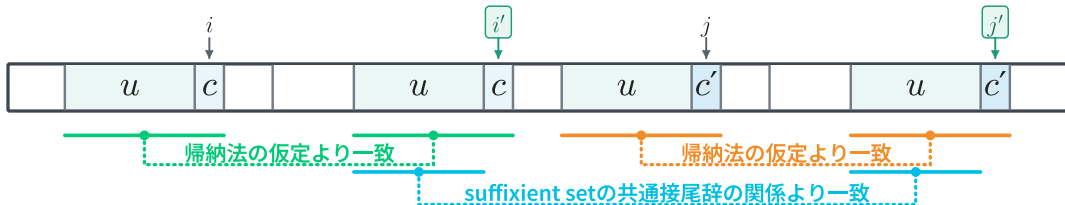


正当性の証明 (4)

19/22

(2) $c \neq c'$ の場合 (続き) :

1. suffixient set の位置 i', j' で終わる uc, uc' の出現について、
 $w[i' - \ell..i' - 1] = w[j' - \ell..j' - 1] = u$
2. 下図の3つの一致関係より、 u の出現同士の一一致関係が $\Lambda(w)$ のみからわかる



目次

1. 部分文字列方程式 (SES)
2. Suffixient set サイズの SES の構成
3. SES・BMS のサイズ関係

- $s(w)$: 文字列 w を表す SES の制約数の最小値
- $b(w)$: 文字列 w を表す BMS のフレーズ数の最小値

定理 2: 最小 SES のサイズ関係

任意の文字列 w に対して $s(w) \leq \min\{b(w), \chi(w) + \sigma(w) - 1\}$ が成り立つ。

- $s(w)$: 文字列 w を表す SES の制約数の最小値
- $b(w)$: 文字列 w を表す BMS のフレーズ数の最小値

定理 2: 最小 SES のサイズ関係

任意の文字列 w に対して $s(w) \leq \min\{b(w), \chi(w) + \sigma(w) - 1\}$ が成り立つ。

事実: $\chi(w) \in \Omega(s(w) \log |w|)$ となる文字列の系列が存在する

- $s(w)$: 文字列 w を表す SES の制約数の最小値
- $b(w)$: 文字列 w を表す BMS のフレーズ数の最小値

定理 2: 最小 SES のサイズ関係

任意の文字列 w に対して $s(w) \leq \min\{b(w), \chi(w) + \sigma(w) - 1\}$ が成り立つ。

事実: $\chi(w) \in \Omega(s(w) \log |w|)$ となる文字列の系列が存在する

自然な疑問: $s(w) \in o(b(w))$ となる文字列の系列は存在するのか？

定理 3: 最小 SES のサイズ

任意の文字列 w に対して $b(w) \leq 4s(w)$ が成り立つ。

→ 任意の文字列に対して、SES と BMS の最小サイズは高々定数倍しか変わらない！

定理 3: 最小 SES のサイズ

任意の文字列 w に対して $b(w) \leq 4s(w)$ が成り立つ。

→ 任意の文字列に対して、SES と BMS の最小サイズは高々定数倍しか変わらない！

定理 4: 最小 SES のサイズ (定理 2,3 から導ける結果)

任意の文字列 w に対して、 $b(w) \in O(\chi(w))$ が成り立つ。

- 部分文字列方程式 (SES) による文字列表現を提案
 - 既存の BMS と同等の圧縮能力を持ち、より自然 (主観)
- SES を利用し、最小 suffixient set サイズ $\chi(w)$ に対して線形サイズの文字列表現を提案
 - このサイズで任意の文字列を表現できるかは未解決問題だった
 - 他の結果との組み合わせにより BMS/suffixient set のサイズ関係も導ける