圧縮指標 : 計算と特定

東京医科歯科大学 の クップル

圧縮指標

最小の

- LZ77, LZ78, LZ-End, lexparse, etc.
- run-length of BWT r
- optimal LZ78
- bidirectional macro scheme b
- 文脈自由文法 g
- 文字列アトラクタ γ
- collage system
- optimal LZ-End

圧縮指標

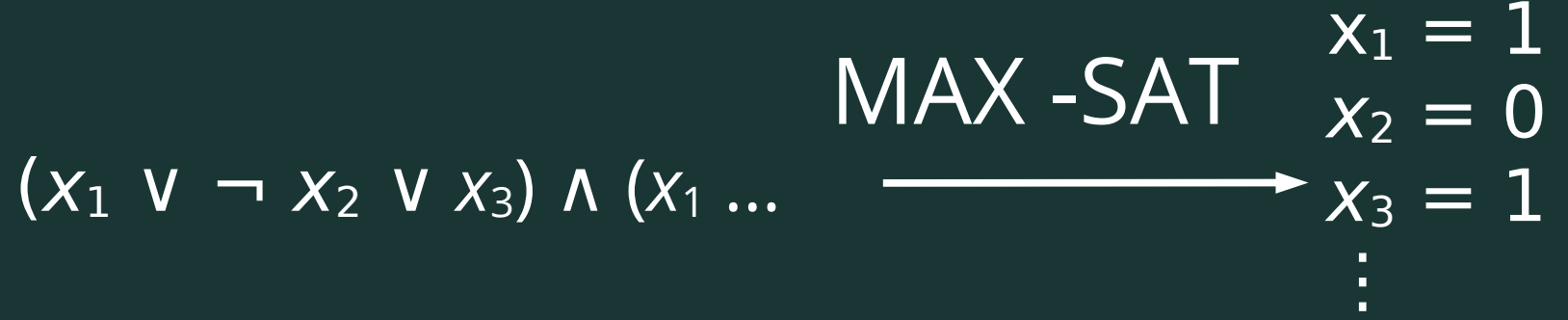
- LZ77, LZ78, LZ-End, lexparse, etc.
- run-length of BWT r
- optimal LZ78
- bidirectional macro scheme b
- 文脈自由文法 g
- 文字列アトラクタ γ
- collage system
- optimal LZ-End

線形時間で計算可能

最小のサイズを求めることは NP 困難

難しそう

圧縮指標の計算



圧縮指標

最小の

- LZ77, LZ78, LZ-End, lexparse, etc.
- run-length of BWT r
- optimal LZ78
- bidirectional macro scheme b
- 文脈自由文法 g
- 文字列アトラクタ γ
- collage system
- optimal LZ-End

線形時間で計算可能

[坂内、後藤、石畠、神田、K、西本 : FPAI 22]: SAT ソルバで高速に計算する手法を提案

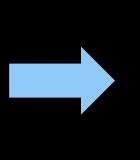
SAT の問題

- 入力:　論理式 CNF
- 出力:　真にする割当

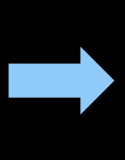
$x_1 \land (x_2 \lor \neg x_3) \land (x_2 \lor x_4 \lor \ldots$

節

$x_1 \land (x_2 \lor \neg x_3)$
$\land (x_2 \lor x_4 \lor \ldots$



SAT ソルバ



$x_1 = 1$
$x_2 = 0$
$x_3 = 1$
$\vdots$

MAX-SAT 問題

SAT を拡張された最大化問題

- hard 節 : 絶対に充足しないといけない節
- soft 節: できるだけ充足

  ⇒ 充足される soft 節の個数を最大化

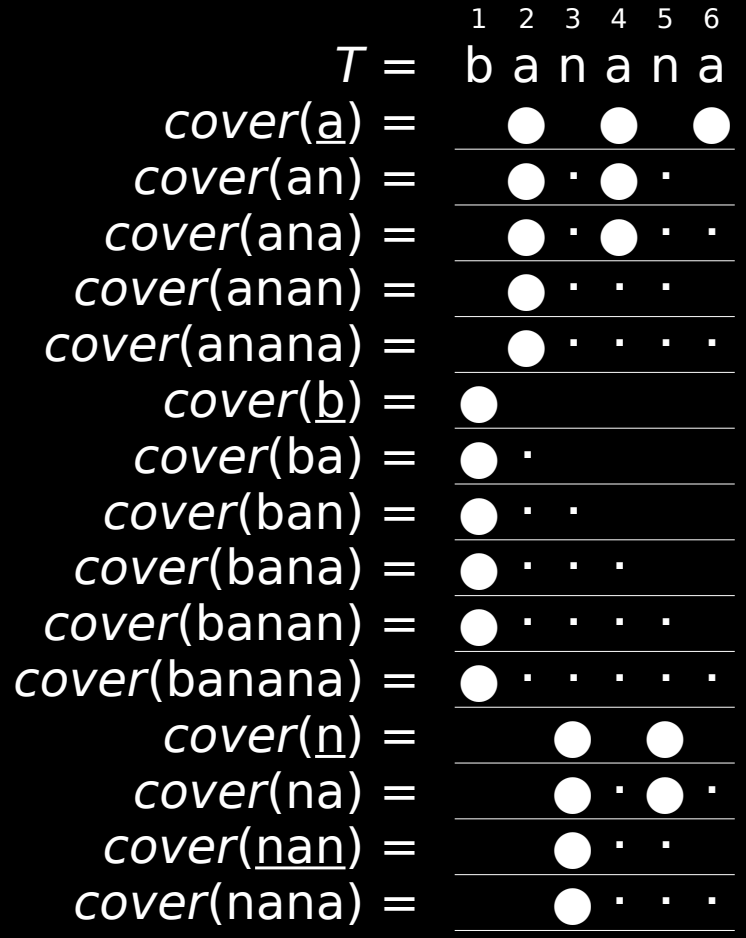
soft 節

- 入力：文字列 T
- 目的:　圧縮指標（g, γ, b など）の最小の個数を定めたい

個数を soft 節で表現

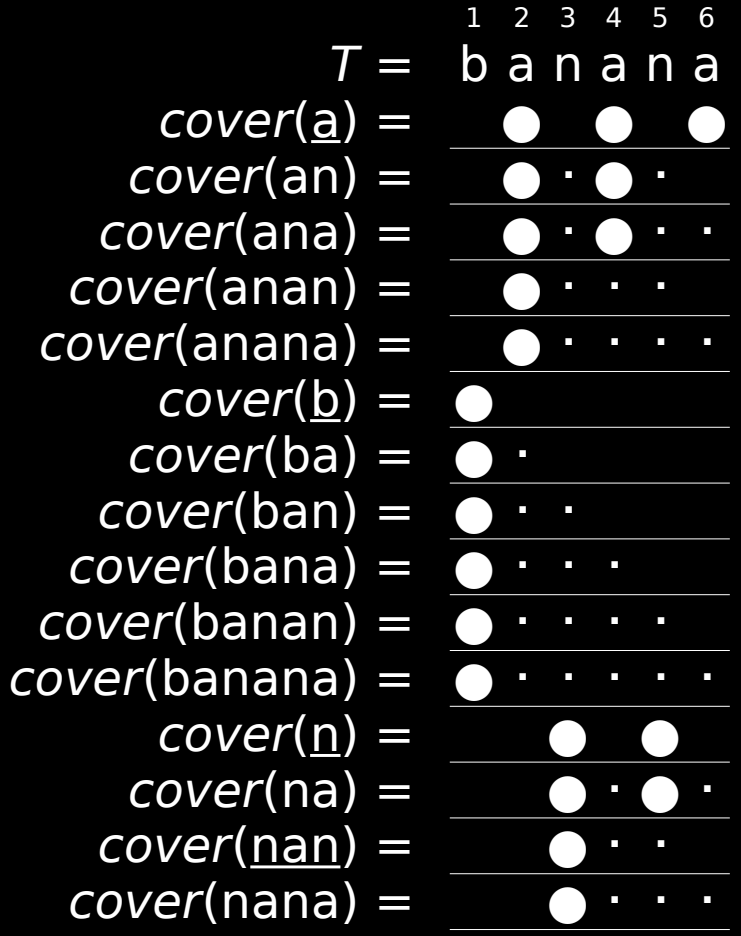
- $T[1..n]$ の位置を個数に割り合って、soft 節を作成
  - $\forall\, i \in [1..n]$: $\neg p_i$
  - つまり、$\Sigma_i\, p_i$ を最小化

文字列アトラクタ



- *cover*(*S*): *T* の中に出現する部分文字列 *S* の出現範囲の位置集合
- 位置 *p* について *p* ∈ *cover*(*S*) のとき、*p* は *S* の出現を串刺しにすると呼ぶ
- 例:位置 6 は a, ana, ... の出現を串刺しにする
- 「·」は cover(*S*) を表現する
- 「●」は *S* の出現開始位置

文字列アトラクタ



- $T[1..n]$ のアトラクタは 以下を満たす T の位置集合 $\Gamma \subseteq [1..n]$
  - 任意の部分文字列 S に対して cover(S) ∩ Γ ≠ ∅
  - ようは、 S の出現を串刺しにする Γ の要素が存在
- すべての位置集合 [1 .. n] は自明なアトラクタ（なおかつ最大の）
- 3 は最小のアトラクタの個数
- Γ = {1,2,3}

hard 節の表現

- $p_i = 1 \Leftrightarrow i \in \Gamma$
- T の任意の部分文字列 S に対して、
  - hard 節 $C_S = \bigvee_{i \in \mathrm{cover}(S)} p_i$ を作成する（アトラクタの制限）
- hard 節を全て満たしたら、$\{i \mid p_i = 1\}$ はアトラクタになる
- なおかつ、充足する soft 節を最大化する p_i の割当について $\{i \mid p_i = 1\}$ は最小アトラクタ
- hard 節の個数は $O(n^2)$
- 節のサイズは $O(n)$

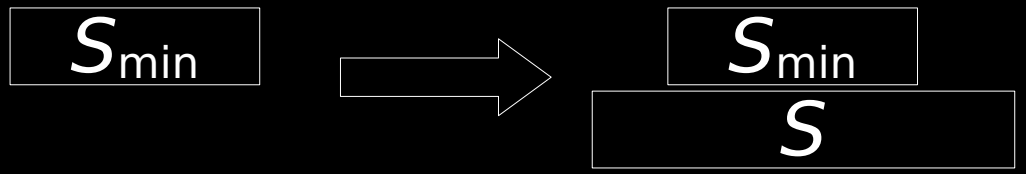
minimal substrings

T の minimal substring は以下の状況を満たす T の部分文字列 S:

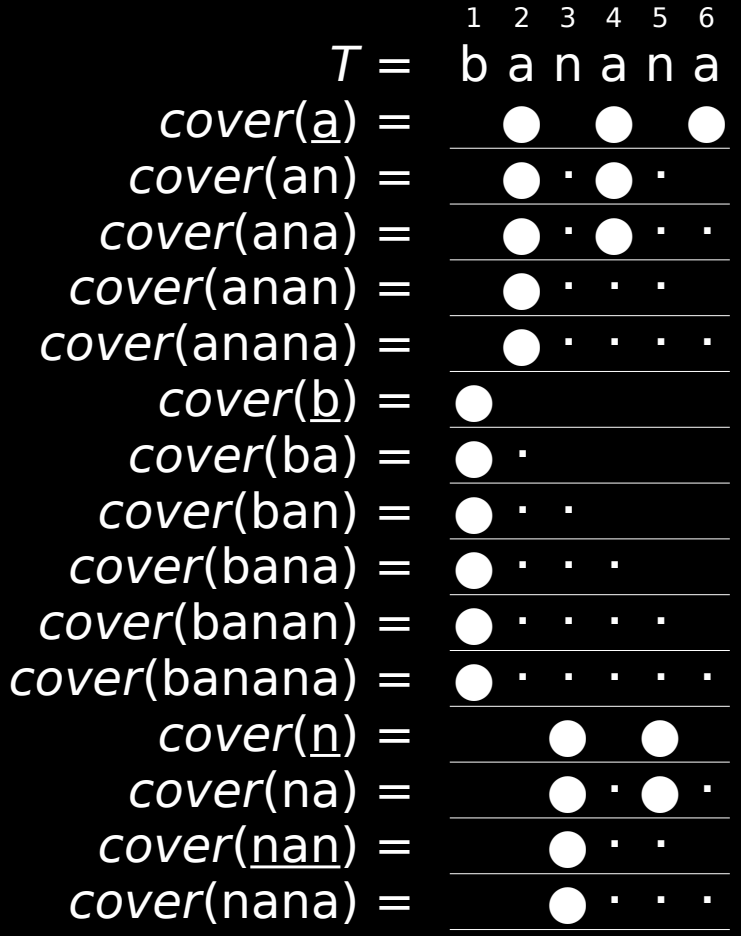
- すべての S の真の部分文字列の出現は S の出現より多い

主張: 任意部分文字列 S は minimal ではないと hard 節 C_S を作成しなくてもよい:

- S は minimal ではないと、S と同じ出現の個数が持つ S の minimal substring $S_{\min}$ が存在がある
- hard 節 $C_{S\min}$ によって、計算したアトラクター の要素 は $S_{\min}$ の出現を串刺しにする ⇒ あの $S_{\min}$ の出現を S の出現に伸ばすことができる



文字列アトラクタ



下線の部分文字列は minimal substrings

- m は minimal substring の個数とすると、
- hard 節の個数は $O(nm)$
- $m = o(n)$ を満たす文字列集合が存在する

問題

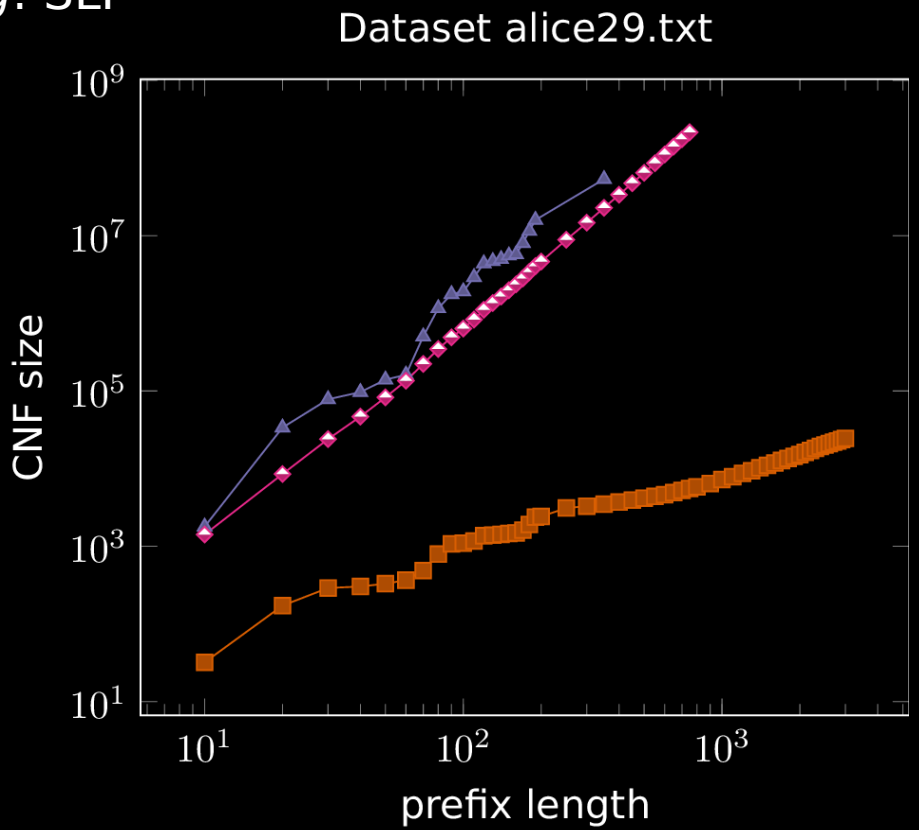
- hard 節をもっと減らすことができる？
- 節のサイズを減らすことができる？
- 個数の下界がある？
- MAX-SAT ソルバで高速に計算可能？

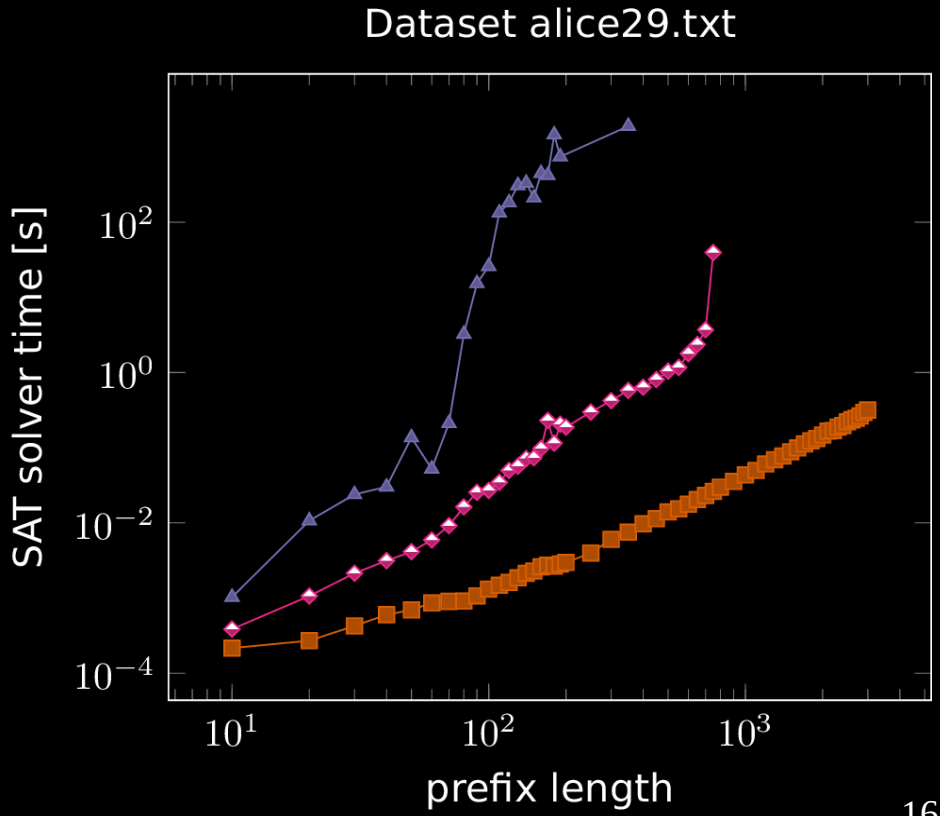
CNF の指標

	アトラクター γ	SLP g	BMS b
Boolean 変数の個数	n	$O(n^3)$	$O(n^3)$
hard 節の集合	$O(nm)$	$O(n^4)$	$O(n^4)$
節のサイズ	$O(n)$	$O(n)$	$O(n^2)$

実験

γ: アトラクター
b: BMS
g: SLP





MAX-SAT について未決問題

- よい近似アルゴリズムは存在がある？ (simulated annealing など)

まだ MAX-SAT の CNF で表現しない問題

- 連超圧縮 SLP
- collage system
- optimal LZ78

解かれたということを小耳に挟んだ

collage system と optimal LZ-End は NP hard かどうか、まだ分からない

圧縮指標の特定
2 uniform morphism

φ(a) = ab
φ(b) = ba

$T_k = \varphi^k(a)$

T_0 = a
T_1 = ab
T_2 = abba
T_3 = abbabaab
⋮

string morphism

三重野、堀山、稲永：
Fibonacci 列の最小文法
を特定できた

特徴な文字列集合に対して、圧縮指標を定めることができる？

- string morphism は以下の状況を満たす写像 φ
  - 任意 $c \in \Sigma$ に対して、$\varphi(c) \in \Sigma^+$
- $S_k = \varphi^k(a) = \varphi(\ldots(\varphi(a))\ldots) \in \Sigma^+$

2-uniform morphism: 任意 $c \in \Sigma$ に対して、$|\varphi(c)| = 2$

- 例: φ(a) = ab, φ(b) = ba ⇒ $T_k = \varphi^k(a)$ k 番目の Thue-Morse 列
- φ(a) = ab, φ(b) = aa ⇒ $P_k = \varphi^k(a)$ k 番目の Period-Doubling 列
- φ をペアー (φ(a),φ(b)) = (ab,ba) で表現できる

2-uniform morphism, σ = 2

$\varphi(a)$	$\varphi(b)$	$\varphi^{(k)}(a)$	γ	z	z_{no}	g
aa	*	a^{2^k}	1	2	$k+1$	$k+1$
ab	ab	$(ab)^{2^k-1}$	2	3	$k+1$	$k+2$
	ba	T_k	4	$2k$	$2k$	$2k+1$
	aa	P_k	2	$2k$	$2k$	$2k+1$
	bb	ab^{2^k-1}	2	3	$2+k$	addition chain
ba	ab	$\overline{T_k}$			same as T_k	
	ba	$(ba)^k$			same as (ab, ab)	
	bb	$b^{2^k-1}a$			same as (ab, bb)	
	aa	P_k^M	2	$\lfloor 3k/2 \rfloor + 1$?	unknown	same as P_k
bb	ab	$\overline{P_{k-1}}^M \overline{P_{k-1}}^M$	2	$\lfloor 3k/2 \rfloor + 2$?	unknown	$z_{no}(P_{k-1}) + 2$
	ba	$\overline{P_{k-1}P_{k-1}}$	2	$2k-1$?	$2k+2$	$z_{no}(P_{k-1}) + 2$
	aa	$\in \{a^{2^k}, b^{2^k}\}$			same as (aa, aa)	
	bb	b^{2^k}			same as (aa, aa)	

- γ: アトラクター
- z : LZ77 の個数
- z_{no} : 重複なし LZ77
- g : 最小文法のサイズ
- $\overline{a} = b$, $\overline{b} = a$
- $(ab)^M = (ba)^M$

かなり大変

性質

- [Rytter 03, Mieno+ ‘22]: $z \leq z_{no} \leq g - \sigma + 1$
- ab^{2^k-1}の最小文法の個数は addition chain と同値する
  - addition chain (OEIS A003313):
    v^x を v の乗算で計算する演算の最小の個数
  - 難しそうけど、$x = 2^k - 1$ なら、最小の個数を簡単に特定できる?

解決したい問題

- $\sigma > 2$ or k-uniform morphism
  (任意の $c \in \Sigma$ に対して、$|\varphi(c)| = k$) に拡張できる?

paper-folding sequence

例: paper-folding sequence: 2-uniform, σ = 4

- φ(a) = ab, φ(b) = cb, φ(c) = ad, φ(d) = cd
- F_0 = a, F_1 = ab, F_2 = abcb, F_3 = ...
- 予想:
  - $g = 4(k - 1)\ \forall\ k \geq 4$
  - γ = 7 ∀ $k \geq 4$（九大の沓掛くん）

沓掛くんの卒論
- $z = 3k - 5$

スライドの由来:　© 後藤啓介さん