r インデックスにおける接尾辞配列を模倣するデータ構造

Christina Boucher [1]

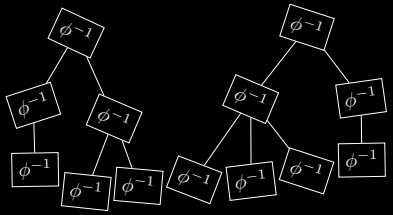
Dominik Köppl [2]

Herman Perera [1]

Massimiliano Rossi [1]

[1]Herbert Wertheim College of Engineering, University of Florida, 米国

[2] 東京医科歯科大学



SPIRE 2022 で発表された結果

今度の話題

目的
r-index 上で、接尾辞配列 (SA) の要素を実際的に速く計算可能？

目的の重要性：

- *r*-index は連超圧縮された FM-index
- *r*-index の SA サンプリングは FM-index より少ない
- 両方の index は検索したパターンの出現を SA で探す

例

入力の文字列

- GATTACAT
- GATACAT
- GATTAGATA

入力を一つの文字列に連結するため：

- 各入力文字列に $ を追加
- # を終端記号として使う

これにより、入力を以下の T に変換できる。

T = GATTACAT$GATACAT$GATTAGATA#

i	SA	順列の行列	BWT
1	27	#GATTACAT$GATACAT$GATTAGAT	A
2	9	$GATACAT$GATTAGATA#GATTACA	T
3	17	$GATTAGATA#GATTACAT$GATACA	T
4	26	A#GATTACAT$GATACAT$GATTAGA	T
5	5	ACAT$GATACAT$GATTAGATA#GAT	T
6	13	ACAT$GATTAGATA#GATTACAT$GA	T
7	22	AGATA#GATTACAT$GATACAT$GAT	T
8	7	AT$GATACAT$GATTAGATA#GATTA	C
9	15	AT$GATTAGATA#GATTACAT$GATA	C
10	24	ATA#GATTACAT$GATACAT$GATTA	G
11	11	ATACAT$GATTAGATA#GATTACAT$	G
12	2	ATTACAT$GATACAT$GATTAGATA#	G
13	19	ATTAGATA#GATTACAT$GATACAT$	G
14	6	CAT$GATACAT$GATTAGATA#GATT	A
15	14	CAT$GATTAGATA#GATTACAT$GAT	A
16	23	GATA#GATTACAT$GATACAT$GATT	A
17	10	GATACAT$GATTAGATA#GATTACAT	$
18	1	GATTACAT$GATACAT$GATTAGATA	#
19	18	GATTAGATA#GATTACAT$GATACAT	$
20	8	T$GATACAT$GATTAGATA#GATTAC	A
21	16	T$GATTAGATA#GATTACAT$GATAC	A
22	25	TA#GATTACAT$GATACAT$GATTAG	A
23	4	TACAT$GATACAT$GATTAGATA#GA	T
24	12	TACAT$GATTAGATA#GATTACAT$G	A
25	21	TAGATA#GATTACAT$GATACAT$GA	T
26	3	TTACAT$GATACAT$GATTAGATA#G	A
27	20	TTAGATA#GATTACAT$GATACAT$G	A

$T :=$ GATTACAT$GATACAT$GATTAGATA# の FM-index は

- パターン出現の個数を計算可能
- パターン出現を見つけるため、SA を利用
- 領域を節約するため、SA をサンプリングする

$T :=$ GATTACAT$GATACAT$GATTAGATA# の FM-index は

- パターン出現の個数を計算可能
- パターン出現を見つけるため、SA を利用
- 領域を節約するため、SA をサンプリングする

r-index は

- BWT の連の境目のみ SA をサンプリングする

i	SA	順列の行列	BWT
1	27	#GATTACAT$GATACAT$GATTAGAT	A
2	9	$GATACAT$GATTAGATA#GATTACA	T
3	17	$GATTAGATA#GATTACAT$GATACA	T
4	26	A#GATTACAT$GATACAT$GATTAGA	T
5	5	ACAT$GATACAT$GATTAGATA#GAT	T
6	13	ACAT$GATTAGATA#GATTACAT$GA	T
7	22	AGATA#GATTACAT$GATACAT$GAT	T
8	7	AT$GATACAT$GATTAGATA#GATTA	C
9	15	AT$GATTAGATA#GATTACAT$GATA	C
10	24	ATA#GATTACAT$GATACAT$GATTA	G
11	11	ATACAT$GATTAGATA#GATTACAT$	G
12	2	ATTACAT$GATACAT$GATTAGATA#	G
13	19	ATTAGATA#GATTACAT$GATACAT$	G
14	6	CAT$GATACAT$GATTAGATA#GATT	A
15	14	CAT$GATTAGATA#GATTACAT$GAT	A
16	23	GATA#GATTACAT$GATACAT$GATT	A
17	10	GATACAT$GATTAGATA#GATTACAT	$
18	1	GATTACAT$GATACAT$GATTAGATA	#
19	18	GATTAGATA#GATTACAT$GATACAT	$
20	8	T$GATACAT$GATTAGATA#GATTAC	A
21	16	T$GATTAGATA#GATTACAT$GATAC	A
22	25	TA#GATTACAT$GATACAT$GATTAG	A
23	4	TACAT$GATACAT$GATTAGATA#GA	T
24	12	TACAT$GATTAGATA#GATTACAT$G	A
25	21	TAGATA#GATTACAT$GATACAT$GA	T
26	3	TTACAT$GATACAT$GATTAGATA#G	A
27	20	TTAGATA#GATTACAT$GATACAT$G	A

SA アクセス

r-index で、$\mathrm{SA}[i]$ をどう計算するか？

足がかりの補題, Gagie+'18

$\mathrm{BWT}[i] = \mathrm{BWT}[i+1] \Rightarrow$
$\mathrm{SA}[i+1] - \mathrm{SA}[i] = \mathrm{SA}[j+1] - \mathrm{SA}[j]$ for
$\mathrm{SA}[j] := \mathrm{SA}[i] - 1$

例

- $\mathrm{SA}[2] = 9$, $\mathrm{SA}[20] = 8$
- $\mathrm{BWT}[2] = \mathrm{BWT}[3] = T$
- $\mathrm{SA}[3] - \mathrm{SA}[2] = \mathrm{SA}[21] - \mathrm{SA}[20]$.

i	SA	順列の行列	BWT
1	27	#GATTACAT$GATACAT$GATTAGAT	A
2	9	$GATACAT$GATTAGATA#GATTACA	T
3	17	$GATTAGATA#GATTACAT$GATACA	T
4	26	A#GATTACAT$GATACAT$GATTAGA	T
5	5	ACAT$GATACAT$GATTAGATA#GAT	T
6	13	ACAT$GATTAGATA#GATTACAT$GA	T
7	22	AGATA#GATTACAT$GATACAT$GAT	T
8	7	AT$GATACAT$GATTAGATA#GATTA	C
9	15	AT$GATTAGATA#GATTACAT$GATA	C
10	24	ATA#GATTACAT$GATACAT$GATTA	G
11	11	ATACAT$GATTAGATA#GATTACAT$	G
12	2	ATTACAT$GATACAT$GATTAGATA#	G
13	19	ATTAGATA#GATTACAT$GATACAT$	G
14	6	CAT$GATACAT$GATTAGATA#GATT	A
15	14	CAT$GATTAGATA#GATTACAT$GAT	A
16	23	GATA#GATTACAT$GATACAT$GATT	A
17	10	GATACAT$GATTAGATA#GATTACAT	$
18	1	GATTACAT$GATACAT$GATTAGATA	#
19	18	GATTAGATA#GATTACAT$GATACAT	$
20	8	T$GATACAT$GATTAGATA#GATTAC	A
21	16	T$GATTAGATA#GATTACAT$GATAC	A
22	25	TA#GATTACAT$GATACAT$GATTAG	A
23	4	TACAT$GATACAT$GATTAGATA#GA	T
24	12	TACAT$GATTAGATA#GATTACAT$G	A
25	21	TAGATA#GATTACAT$GATACAT$GA	T
26	3	TTACAT$GATACAT$GATTAGATA#G	A
27	20	TTAGATA#GATTACAT$GATACAT$G	A

系

- $k \geq 0$ は以下の状況を満たす整数とする。すべて $\mathrm{SA}[i'] \in [\mathrm{SA}[i] - k + 1..\mathrm{SA}[i]]$ を満たす i に対して、$\mathrm{BWT}[i'] = \mathrm{BWT}[i' + 1]$
- $\mathrm{SA}[j] := \mathrm{SA}[i] - k$ を設定する。

従って、

- $\mathrm{BWT}[j] \neq \mathrm{BWT}[j + 1]$ ただし
- $\mathrm{SA}[i + 1] - \mathrm{SA}[i] = \mathrm{SA}[j + 1] - \mathrm{SA}[j]$.

i	SA	順列の行列	BWT
1	27	#GATTACAT$GATACAT$GATTAGAT	A
2	9	$GATACAT$GATTAGATA#GATTACA	T
3	17	$GATTAGATA#GATTACAT$GATACA	T
4	26	A#GATTACAT$GATACAT$GATTAGA	T
5	5	ACAT$GATACAT$GATTAGATA#GAT	T
6	13	ACAT$GATTAGATA#GATTACAT$GA	T
7	22	AGATA#GATTACAT$GATACAT$GAT	T
8	7	AT$GATACAT$GATTAGATA#GATTA	C
9	15	AT$GATTAGATA#GATTACAT$GATA	C
10	24	ATA#GATTACAT$GATACAT$GATTA	G
11	11	ATACAT$GATTAGATA#GATTACAT$	G
12	2	ATTACAT$GATACAT$GATTAGATA#	G
13	19	ATTAGATA#GATTACAT$GATACAT$	G
14	6	CAT$GATACAT$GATTAGATA#GATT	A
15	14	CAT$GATTAGATA#GATTACAT$GAT	A
16	23	GATA#GATTACAT$GATACAT$GATT	A
17	10	GATACAT$GATTAGATA#GATTACAT	$
18	1	GATTACAT$GATACAT$GATTAGATA	#
19	18	GATTAGATA#GATTACAT$GATACAT	$
20	8	T$GATACAT$GATTAGATA#GATTAC	A
21	16	T$GATTAGATA#GATTACAT$GATAC	A
22	25	TA#GATTACAT$GATACAT$GATTAG	A
23	4	TACAT$GATACAT$GATTAGATA#GA	T
24	12	TACAT$GATTAGATA#GATTACAT$G	A
25	21	TAGATA#GATTACAT$GATACAT$GA	T
26	3	TTACAT$GATACAT$GATTAGATA#G	A
27	20	TTAGATA#GATTACAT$GATACAT$G	A

系

- $k \geq 0$ は以下の状況を満たす整数とする。すべて $\mathrm{SA}[i'] \in [\mathrm{SA}[i] - k + 1..\mathrm{SA}[i]]$ を満たす i に対して、$\mathrm{BWT}[i'] = \mathrm{BWT}[i' + 1]$
- $\mathrm{SA}[j] := \mathrm{SA}[i] - k$ を設定する。

従って、

- $\mathrm{BWT}[j] \neq \mathrm{BWT}[j + 1]$ ただし
- $\mathrm{SA}[i + 1] - \mathrm{SA}[i] = \mathrm{SA}[j + 1] - \mathrm{SA}[j]$.

i	SA	順列の行列	BWT
1	27	#GATTACAT$GATACAT$GATTAGAT	A
2	9	$GATACAT$GATTAGATA#GATTACA	T
3	17	$GATTAGATA#GATTACAT$GATACA	T
4	26	A#GATTACAT$GATACAT$GATTAGA	T
5	5	ACAT$GATACAT$GATTAGATA#GAT	T
6	13	ACAT$GATTAGATA#GATTACAT$GA	T
7	22	AGATA#GATTACAT$GATACAT$GAT	T
8	7	AT$GATACAT$GATTAGATA#GATTA	C
9	15	AT$GATTAGATA#GATTACAT$GATA	C
10	24	ATA#GATTACAT$GATACAT$GATTA	G
11	11	ATACAT$GATTAGATA#GATTACAT$	G
12	2	ATTACAT$GATACAT$GATTAGATA#	G
13	19	ATTAGATA#GATTACAT$GATACAT$	G
14	6	CAT$GATACAT$GATTAGATA#GATT	A
15	14	CAT$GATTAGATA#GATTACAT$GAT	A
16	23	GATA#GATTACAT$GATACAT$GATT	A
17	10	GATACAT$GATTAGATA#GATTACAT	$
18	1	GATTACAT$GATACAT$GATTAGATA	#
19	18	GATTAGATA#GATTACAT$GATACAT	$
20	8	T$GATACAT$GATTAGATA#GATTAC	A
21	16	T$GATTAGATA#GATTACAT$GATAC	A
22	25	TA#GATTACAT$GATACAT$GATTAG	A
23	4	TACAT$GATACAT$GATTAGATA#GA	T
24	12	TACAT$GATTAGATA#GATTACAT$G	A
25	21	TAGATA#GATTACAT$GATACAT$GA	T
26	3	TTACAT$GATACAT$GATTAGATA#G	A
27	20	TTAGATA#GATTACAT$GATACAT$G	A

系

- $k \geq 0$ は以下の状況を満たす整数とする。すべて $\mathrm{SA}[i'] \in [\mathrm{SA}[i] - k + 1..\mathrm{SA}[i]]$ を満たす i に対して、$\mathrm{BWT}[i'] = \mathrm{BWT}[i' + 1]$
- $\mathrm{SA}[j] := \mathrm{SA}[i] - k$ を設定する。

従って、

- $\mathrm{BWT}[j] \neq \mathrm{BWT}[j + 1]$ ただし
- $\mathrm{SA}[i + 1] - \mathrm{SA}[i] = \mathrm{SA}[j + 1] - \mathrm{SA}[j]$.

i	SA	順列の行列	BWT
1	27	#GATTACAT$GATACAT$GATTAGAT	A
2	9	$GATACAT$GATTAGATA#GATTACA	T
3	17	$GATTAGATA#GATTACAT$GATACA	T
4	26	A#GATTACAT$GATACAT$GATTAGA	T
5	5	ACAT$GATACAT$GATTAGATA#GAT	T
6	13	ACAT$GATTAGATA#GATTACAT$GA	T
7	22	AGATA#GATTACAT$GATACAT$GAT	T
8	7	AT$GATACAT$GATTAGATA#GATTA	C
9	15	AT$GATTAGATA#GATTACAT$GATA	C
10	24	ATA#GATTACAT$GATACAT$GATTA	G
11	11	ATACAT$GATTAGATA#GATTACAT$	G
12	2	ATTACAT$GATACAT$GATTAGATA#	G
13	19	ATTAGATA#GATTACAT$GATACAT$	G
14	6	CAT$GATACAT$GATTAGATA#GATT	A
15	14	CAT$GATTAGATA#GATTACAT$GAT	A
16	23	GATA#GATTACAT$GATACAT$GATT	A
17	10	GATACAT$GATTAGATA#GATTACAT	$
18	1	GATTACAT$GATACAT$GATTAGATA	#
19	18	GATTAGATA#GATTACAT$GATACAT	$
20	8	T$GATACAT$GATTAGATA#GATTAC	A
21	16	T$GATTAGATA#GATTACAT$GATAC	A
22	25	TA#GATTACAT$GATACAT$GATTAG	A
23	4	TACAT$GATACAT$GATTAGATA#GA	T
24	12	TACAT$GATTAGATA#GATTACAT$G	A
25	21	TAGATA#GATTACAT$GATACAT$GA	T
26	3	TTACAT$GATACAT$GATTAGATA#G	A
27	20	TTAGATA#GATTACAT$GATACAT$G	A

系

- $k \geq 0$ は以下の状況を満たす整数とする。すべて $\mathrm{SA}[i'] \in [\mathrm{SA}[i] - k + 1..\mathrm{SA}[i]]$ を満たす i に対して、$\mathrm{BWT}[i'] = \mathrm{BWT}[i' + 1]$
- $\mathrm{SA}[j] := \mathrm{SA}[i] - k$ を設定する。

従って、

- $\mathrm{BWT}[j] \neq \mathrm{BWT}[j + 1]$ ただし
- $\mathrm{SA}[i + 1] - \mathrm{SA}[i] = \mathrm{SA}[j + 1] - \mathrm{SA}[j]$.

i	SA	順列の行列	BWT
1	27	#GATTACAT$GATACAT$GATTAGAT	A
2	9	$GATACAT$GATTAGATA#GATTACA	T
3	17	$GATTAGATA#GATTACAT$GATACA	T
4	26	A#GATTACAT$GATACAT$GATTAGA	T
5	5	ACAT$GATACAT$GATTAGATA#GAT	T
6	13	ACAT$GATTAGATA#GATTACAT$GA	T
7	22	AGATA#GATTACAT$GATACAT$GAT	T
8	7	AT$GATACAT$GATTAGATA#GATTA	C
9	15	AT$GATTAGATA#GATTACAT$GATA	C
10	24	ATA#GATTACAT$GATACAT$GATTA	G
11	11	ATACAT$GATTAGATA#GATTACAT$	G
12	2	ATTACAT$GATACAT$GATTAGATA#	G
13	19	ATTAGATA#GATTACAT$GATACAT$	G
14	6	CAT$GATACAT$GATTAGATA#GATT	A
15	14	CAT$GATTAGATA#GATTACAT$GAT	A
16	23	GATA#GATTACAT$GATACAT$GATT	A
17	10	GATACAT$GATTAGATA#GATTACAT	$
18	1	GATTACAT$GATACAT$GATTAGATA	#
19	18	GATTAGATA#GATTACAT$GATACAT	$
20	8	T$GATACAT$GATTAGATA#GATTAC	A
21	16	T$GATTAGATA#GATTACAT$GATAC	A
22	25	TA#GATTACAT$GATACAT$GATTAG	A
23	4	TACAT$GATACAT$GATTAGATA#GA	T
24	12	TACAT$GATTAGATA#GATTACAT$G	A
25	21	TAGATA#GATTACAT$GATACAT$GA	T
26	3	TTACAT$GATACAT$GATTAGATA#G	A
27	20	TTAGATA#GATTACAT$GATACAT$G	A

系

- $k \geq 0$ は以下の状況を満たす整数とする。すべて $\mathrm{SA}[i'] \in [\mathrm{SA}[i] - k + 1..\mathrm{SA}[i]]$ を満たす i に対して、$\mathrm{BWT}[i'] = \mathrm{BWT}[i' + 1]$
- $\mathrm{SA}[j] := \mathrm{SA}[i] - k$ を設定する。

従って、

- $\mathrm{BWT}[j] \neq \mathrm{BWT}[j + 1]$ ただし
- $\mathrm{SA}[i + 1] - \mathrm{SA}[i] = \mathrm{SA}[j + 1] - \mathrm{SA}[j]$.

i	SA	順列の行列	BWT
1	27	#GATTACAT$GATACAT$GATTAGAT	A
2	9	$GATACAT$GATTAGATA#GATTACA	T
3	17	$GATTAGATA#GATTACAT$GATACA	T
4	26	A#GATTACAT$GATACAT$GATTAGA	T
5	5	ACAT$GATACAT$GATTAGATA#GAT	T
6	13	ACAT$GATTAGATA#GATTACAT$GA	T
7	22	AGATA#GATTACAT$GATACAT$GAT	T
8	7	AT$GATACAT$GATTAGATA#GATTA	C
9	15	AT$GATTAGATA#GATTACAT$GATA	C
10	24	ATA#GATTACAT$GATACAT$GATTA	G
11	11	ATACAT$GATTAGATA#GATTACAT$	G
12	2	ATTACAT$GATACAT$GATTAGATA#	G
13	19	ATTAGATA#GATTACAT$GATACAT$	G
14	6	CAT$GATACAT$GATTAGATA#GATT	A
15	14	CAT$GATTAGATA#GATTACAT$GAT	A
16	23	GATA#GATTACAT$GATACAT$GATT	A
17	10	GATACAT$GATTAGATA#GATTACAT	$
18	1	GATTACAT$GATACAT$GATTAGATA	#
19	18	GATTAGATA#GATTACAT$GATACAT	$
20	8	T$GATACAT$GATTAGATA#GATTAC	A
21	16	T$GATTAGATA#GATTACAT$GATAC	A
22	25	TA#GATTACAT$GATACAT$GATTAG	A
23	4	TACAT$GATACAT$GATTAGATA#GA	T
24	12	TACAT$GATTAGATA#GATTACAT$G	A
25	21	TAGATA#GATTACAT$GATACAT$GA	T
26	3	TTACAT$GATACAT$GATTAGATA#G	A
27	20	TTAGATA#GATTACAT$GATACAT$G	A

系

- $k \geq 0$ は以下の状況を満たす整数とする。すべて $\mathrm{SA}[i'] \in [\mathrm{SA}[i] - k + 1..\mathrm{SA}[i]]$ を満たす i に対して、 $\mathrm{BWT}[i'] = \mathrm{BWT}[i' + 1]$
- $\mathrm{SA}[j] := \mathrm{SA}[i] - k$ を設定する。

従って、

- $\mathrm{BWT}[j] \neq \mathrm{BWT}[j + 1]$ ただし
- $\mathrm{SA}[i + 1] - \mathrm{SA}[i] = \mathrm{SA}[j + 1] - \mathrm{SA}[j]$.

正確性の理由：
各 backward step でテキスト順序の前の文字に移動する

i	SA	順列の行列	BWT
1	27	#GATTACAT$GATACAT$GATTAGAT	A
2	9	$GATACAT$GATTAGATA#GATTACA	T
3		$GATTAGATA#GATTACAT$GATACA	T
4		A#GATTACAT$GATACAT$GATTAGA	T
5		ACAT$GATACAT$GATTAGATA#GAT	T
6		ACAT$GATTAGATA#GATTACAT$GA	T
7	22	AGATA#GATTACAT$GATACAT$GAT	T
8	7	AT$GATACAT$GATTAGATA#GATTA	C
9	15	AT$GATTAGATA#GATTACAT$GATA	C
10	24	ATA#GATTACAT$GATACAT$GATTA	G
11		ATACAT$GATTAGATA#GATTACAT$	G
12		ATTACAT$GATACAT$GATTAGATA#	G
13	19	ATTAGATA#GATTACAT$GATACAT$	G
14	6	CAT$GATACAT$GATTAGATA#GATT	A
15		CAT$GATTAGATA#GATTACAT$GAT	A
16	23	GATA#GATTACAT$GATACAT$GATT	A
17	10	GATACAT$GATTAGATA#GATTACAT	$
18	1	GATTACAT$GATACAT$GATTAGATA	#
19	18	GATTAGATA#GATTACAT$GATACAT	$
20	8	T$GATACAT$GATTAGATA#GATTAC	A
21		T$GATTAGATA#GATTACAT$GATAC	A
22	25	TA#GATTACAT$GATACAT$GATTAG	A
23	4	TACAT$GATACAT$GATTAGATA#GA	T
24	12	TACAT$GATTAGATA#GATTACAT$G	A
25	21	TAGATA#GATTACAT$GATACAT$GA	T
26	3	TTACAT$GATACAT$GATTAGATA#G	A
27	20	TTAGATA#GATTACAT$GATACAT$G	A

r-index の $\mathcal{O}(r)$ 領域

以下の整数を格納する

- $\mathcal{S}[x]$: x 番目の連の開始位置の SA 値
- $\mathcal{E}[x]$: x 番目の連の終了位置の SA 値

ただし $x \in [1..r]$, r は BWT の連の個数

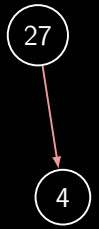
計算に用いるには、$\mathcal{E}$ について以下のクエリが必要：

- $\mathcal{E}.\mathrm{pred}(p) : \max\{q \in \mathcal{E} : q \leq p\}$
- $\mathcal{E}.\mathrm{succ}(p) : \min\{q \in \mathcal{E} : q > p\}$

上記 2 つを利用するために、$\mathcal{E}$ の上で predecessor と successor データ構造を構築

$\mathcal{E}.pred(\mathcal{S}[x+1])$

x		$\mathcal{S}[x]$	$\mathcal{E}[x]$
1	27	27	27
2	4	9	22
3	4	7	15
4	23	24	19
5	4	6	23
6	10	10	10
7	1	1	1
8	18	18	18
9	4	8	25
10	4	4	4
11	12	12	12
12	21	21	21
13	1	3	20

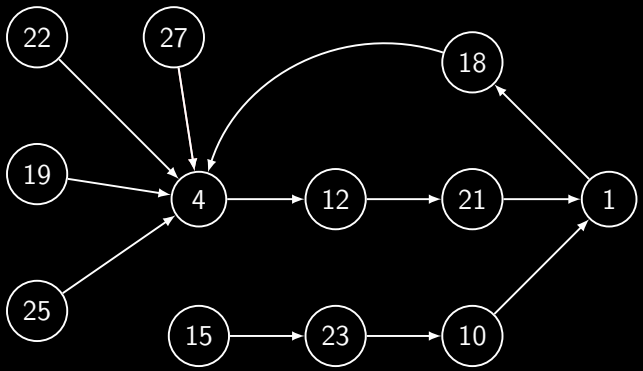


ϕ^{-1} グラフ

- $\mathcal{E}$ の各要素はノードになる
- $\mathcal{E}[x]$ から $\mathcal{E}.\mathrm{pred}(\mathcal{S}[x+1])$ までのアークを書く
  (ソートした $\mathcal{E} =$ $[1, 4, 10, 12, 15, 18, 19, 20, 21, 22, 23, 25, 27]$)

$\mathcal{E}.pred(\mathcal{S}[x+1])$

x		$\mathcal{S}[x]$	$\mathcal{E}[x]$
1		27	27
2	←	4	22
3	←	4	15
4	←	23	19
5	←	4	23
6	←	10	10
7	←	1	1
8	←	18	18
9	←	4	25
10	←	4	4
11	←	12	12
12	←	21	21
13	←	1	20

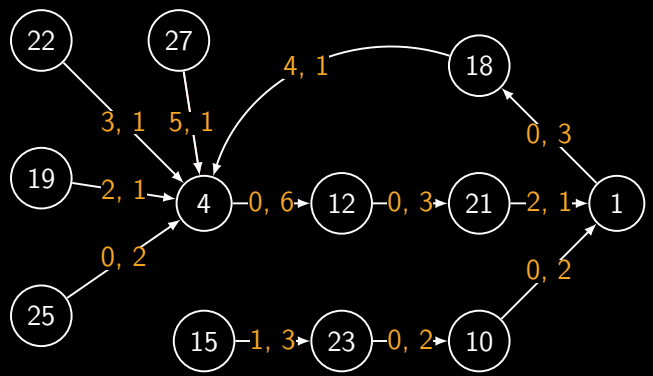


ϕ^{-1} グラフ

- $\mathcal{E}$ の各要素はノードになる
- $\mathcal{E}[x]$ から $\mathcal{E}.\mathrm{pred}(\mathcal{S}[x+1])$ までのアークを書く
  (ソートした $\mathcal{E}=$ $[1,4,10,12,15,18,19,20,21,22,23,25,27]$)

$\mathcal{E}.pred(\mathcal{S}[x+1])$

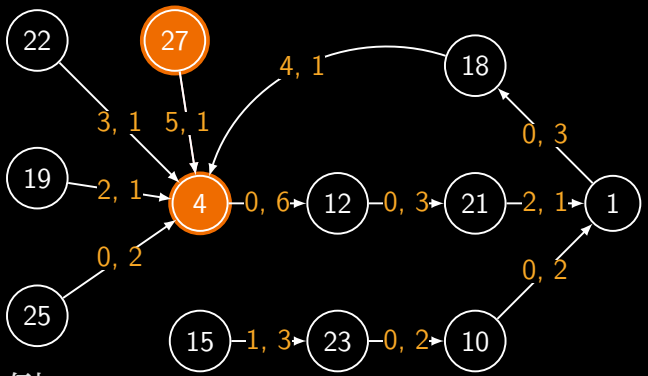
x	/	$\mathcal{S}[x]$	$\mathcal{E}[x]$	c_x	ℓ_x
1	27	27	27	5	1
2	4 ←	9	22	3	1
3	4 ←	7	15	1	3
4	23 ←	24	19	2	1
5	4 ←	6	23	0	2
6	10 ←	10	10	0	2
7	1 ←	1	1	0	3
8	18 ←	18	18	4	1
9	4 ←	8	25	0	2
10	4 ←	4	4	0	6
11	12 ←	12	12	0	3
12	21 ←	21	21	0	2
13	1 ←	3	20	-	2



- x 番目の連のアークにコスト c_x とリミット ℓ_x をラベル付ける、ただし
- $c_x := \mathcal{S}[x+1] - \mathcal{E}.\mathrm{pred}(\mathcal{S}[x+1])$
- $\ell_x := \mathcal{E}.\mathrm{succ}(\mathcal{E}[x]) - \mathcal{E}[x]$
  (ソートした $\mathcal{E} =$
  $[1, 4, 10, 12, 15, 18, 19, 20, 21, 22, 23, 25, 27]$)

SA[i] から SA[$i+1$] の計算

1. 開始ノード $p := \mathcal{E}.\mathrm{pred}(\mathrm{SA}[i])$
2. 最初のコストは $c \leftarrow \mathrm{SA}[i] - p$
3. 足したコスト c はアークのリミットを超えると、止める
4. アークのコストを c に足し、アークで繋いでいるノード v を訪れる
5. SA 値 $v + c$ を出力
6. 3 番に戻って繰り返す

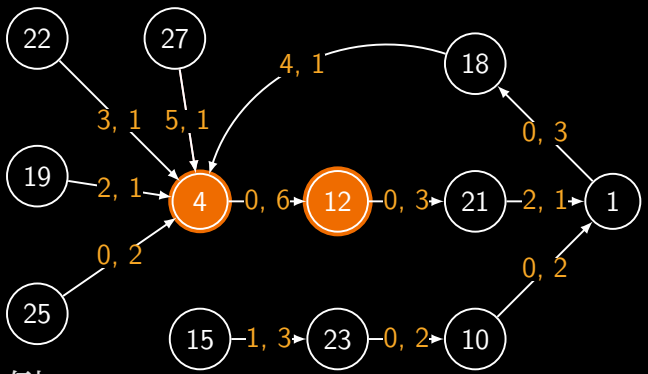


例

- SA[1] $= 27, p = 27, c \leftarrow c_0 = 0$
- アーク $(27, 4)$ のリミットは $1 > c \Rightarrow 4$ を訪れる
- アーク $(27, 4)$ のコスト 5 を $c \leftarrow c_0 + 5 = 5$ に追加
- SA[2] $= 9$ はノード 4 とコスト 5 の和

SA[i] から SA[$i+1$] の計算

1. 開始ノード $p := \mathcal{E}.\mathrm{pred}(\mathrm{SA}[i])$
2. 最初のコストは $c \leftarrow \mathrm{SA}[i] - p$
3. 足したコスト c はアークのリミットを超えると、止める
4. アークのコストを c に足し、アークで繋いでいるノード v を訪れる
5. SA 値 $v + c$ を出力
6. 3 番に戻って繰り返す

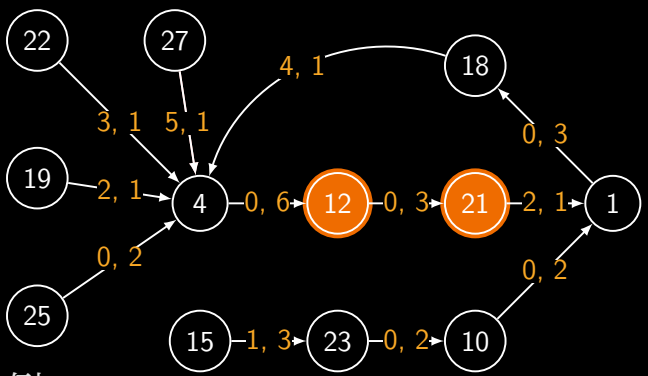


例

- ノード 4 を訪れたとき、コストが $c = 5$ になった
- 次のアーク $(4, 12)$ のリミットは $6 > c \Rightarrow 12$ を訪れる
- $\mathrm{SA}[3] = 17$ はノード 12 とコスト 5 の和

SA[i] から SA[$i+1$] の計算

1. 開始ノード $p := \mathcal{E}.\mathrm{pred}(\mathrm{SA}[i])$
2. 最初のコストは $c \leftarrow \mathrm{SA}[i] - p$
3. 足したコスト c はアークのリミットを超えると、止める
4. アークのコストを c に足し、アークで繋いでいるノード v を訪れる
5. SA 値 $v + c$ を出力
6. 3 番に戻って繰り返す

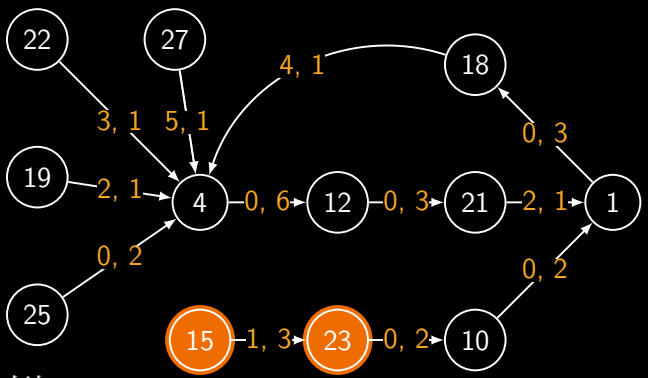


例

- ノード 12 を訪れたとき、コストが $c = 5$ になった
- 次のアーク $(12, 21)$ のリミットは $3 < c \Rightarrow$ 止める
- $\mathrm{SA}[3] = 17$ で続ける
- $p \leftarrow \mathcal{E}.\mathrm{pred}(17) = 15, c_0 \leftarrow 2$

SA[i] から SA[$i+1$] の計算

1. 開始ノード $p := \mathcal{E}.\mathrm{pred}(\mathrm{SA}[i])$
2. 最初のコストは $c \leftarrow \mathrm{SA}[i] - p$
3. 足したコスト c はアークのリミットを超えると、止める
4. アークのコストを c に足し、アークで繋いでいるノード v を訪れる
5. SA 値 $v + c$ を出力
6. 3 番に戻って繰り返す

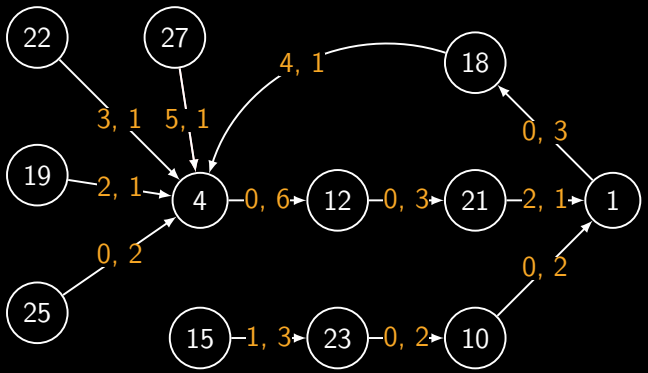


例

- $p \leftarrow \mathcal{E}.\mathrm{pred}(17) = 15, c \leftarrow c_0 = 2$
- 次のアーク $(15, 23)$ のリミットは $3 > c \Rightarrow$ 続ける
- アーク $(15, 23)$ のコスト 1 を $c \leftarrow c_0 + 1 = 3$ に追加
- SA[4] = 26 はノード 23 とコスト 3 の和

SA[i] から SA[$i+1$] の計算

1. 開始ノード $p := \mathcal{E}.\mathrm{pred}(\mathrm{SA}[i])$
2. 最初のコストは $c \leftarrow \mathrm{SA}[i] - p$
3. 足したコスト c はアークのリミットを超えると、止める
4. アークのコストを c に足し、アークで繋いでいるノード v を訪れる
5. SA 値 $v + c$ を出力
6. 3 番に戻って繰り返す



高速の希望

- 各渡ったアークに対して、 $\mathcal{E}$ の predecessor クエリを省いた
- 辿った距離が長い場合でも，predecessor は 1 回だけで十分

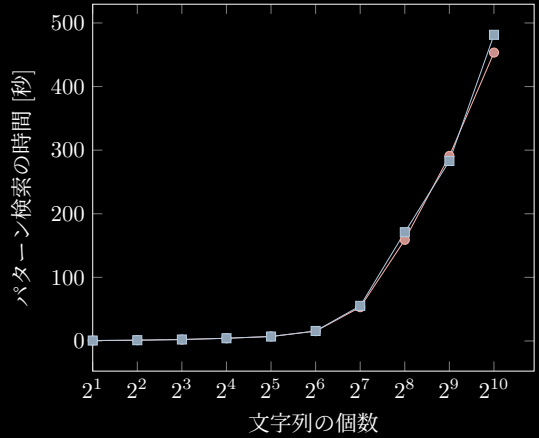
実験

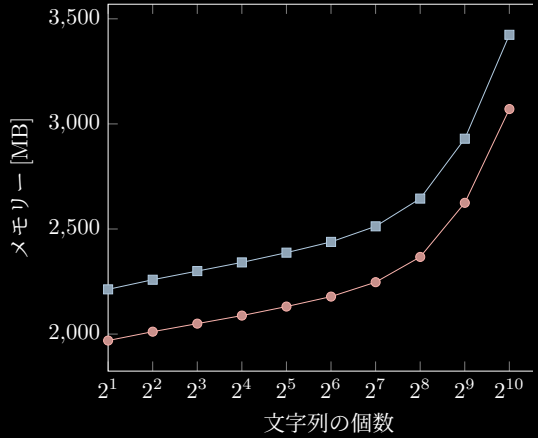
データセット

- 1000 Genomes Project からのヒト１９番目の染色体の遺伝子データ

実験環境

- AMD EPYC 75F3 32-core processor
- 512 GB ラム
- 64-bit Linux





19番目の染色体

r-index ϕ^{-1} グラフ

ϕ^{-1} グラフはメリットがなさそう

まとめ

ϕ^{-1} グラフを紹介した

- r インデックスのように $\mathcal{O}(r)$ 領域を取る
- SA アクセスを提供できる
- クエリの時間は以下の 2 つに比例する
  - 連の長さ
  - コストとリミットの値

未解決問題

- 必要な predecessor クエリの個数に関する理論的な解析

ご清聴ありがとうございました。