FM-Indexing Grammars Induced by Suffix Sorting for Long Patterns

Jin Jie Deng (Jamie), Wing-Kai Hon, Dominik Köppl, Kunihiko Sadakane

Preprint is on arXiv: https://arxiv.org/abs/2110.01181

Outline

- Text Index
- FM-Index
- Geometric-BWT
- Our Index
- Experimental Result
- Future Work

Text Index

- Given a text T over alphabet Σ, a text index is a data structure built from T supporting the following queries with pattern P.
- exist(P): Does P occur in T?
- **count(P): How often does P occur in T?**
- locate(P): Where does P occur in T?

      0123456789012
T = bananabanana$
P = ana

Occurrences:
bananabanana$
bananabanana$
bananabanana$
bananabanana$

exist(P) = true
count(P) = 4
locate(P) = (1, 3, 7, 9)

FM-Index [Ferragina & Manzini JACM‘05]

T = bananabanana$

n rotations

bananabanana$
ananabanana$b
nanabanana$ba
anabanana$ban
nabanana$bana
abanana$banan
banana$banana
anana$bananab
nana$bananaba
ana$bananaban
na$bananabana
a$bananabanan
$bananabanana

sort

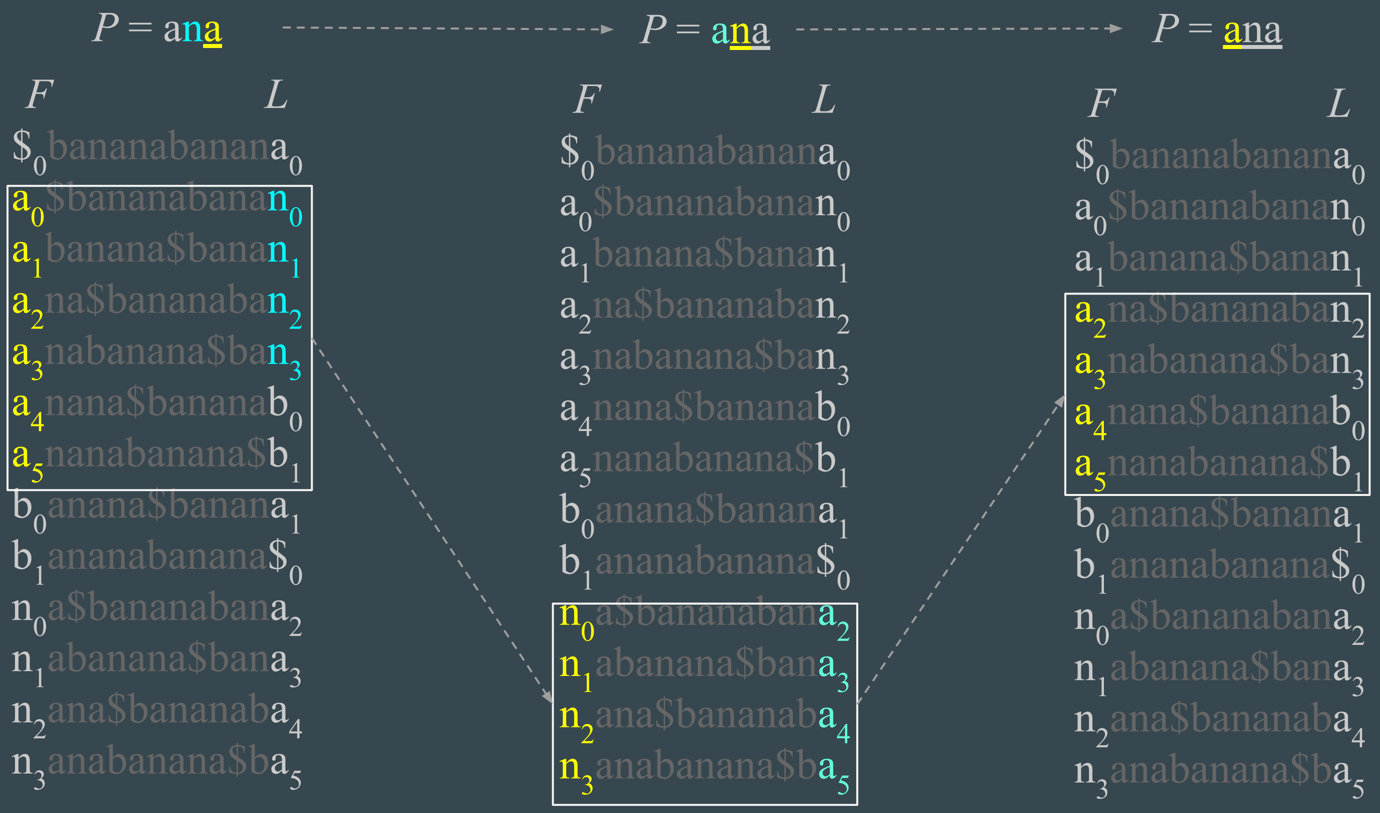
$bananabanana
a$bananabanan
abanana$banan
ana$bananaban
anabanana$ban
anana$bananab
ananabanana$b
banana$banana
bananabanana$
na$bananabana
nabanana$bana
nana$bananaba
nanabanana$ba

store only information of first (F) & last (L) columns

F		L
$	bananabanan	a
a	$bananabana	n
a	banana$bana	n
a	na$bananaba	n
a	nabanana$ba	n
a	nana$banana	b
a	nanabanana$	b
b	anana$banan	a
b	ananabanana	$
n	a$bananaban	a
n	abanana$ban	a
n	ana$bananab	a
n	anabanana$b	a

L : Burrows-Wheeler Transform (BWT) [Burrows & Wheeler TR‘94]

FM-Index (Counting by Backward-Searching)



FM-Index (LF-Mapping)

- begin(F, c): where is begin of maximal-character-run of character c ?
- rank(L, i, c): How many character c is in prefix $L[0 : i\text{-}1]$ of string L ?
- LF-Mapping(i): begin(F, c) + rank(L, i, c)
- Counting P on FM-index is reduced to O($|P|$) LF-Mappings.
- begin(F, c) can be supported by simple prefix-sum array.
- rank(L, i, c) can be support by Wavelet Tree [Grossi et al, SODA'03]
- Disadvantage: LF-Mapping is not cache-friendly.

Geometric BWT [Yu-Feng Cieng et al. Algorithmica'05]

- λ-factorization: Viewing λ consecutive characters as single meta-character (fitted into single machine word).
- $T^{(\lambda)}$: T follows λ-factorization
- $P^{(\lambda,\, k)}$: P follows λ-factorization and the last factor is of length k, $1 \le k \le \lambda$
- Reducing pattern counting to backward searching + 4-sided range counting.

$T =$ bananabanana\$
$P =$ anana

$\xrightarrow{\lambda = 2}$

$T^{(2)} =$ ba.na.na.ba.na.na.\$\$
$P^{(2,\, 1)} =$ an.an.a?
$P^{(2,\, 2)} =$?a.na.na

Geometric BWT

$T^{(2)} = \text{ba.na.na.ba.na.na.\$\$}$

colexgraphical rank range

$P^{(2,1)} = \text{an.an.}\underline{\text{a?}}$

$F^{(2)}$	$L^{(2)}$
$\$\$_0$	na_0
ba_0	na_1
ba_1	$\$\$_0$
na_0	na_2
na_1	na_3
na_2	ba_0
na_3	ba_1

There is no <u>a?</u> in $F^{(2)}$

$P^{(2,2)} = \text{?a.na.}\underline{\text{na}}$ → $P^{(2,2)} = \underline{\text{?a}}.\underline{\text{na}}.\underline{\text{na}}$ → $P^{(2,2)} = \underline{\text{?a}}.\underline{\text{na}}.\underline{\text{na}}$

$F^{(2)}$	$L^{(2)}$
$\$\$_0$	na_0
ba_0	na_1
ba_1	$\$\$_0$
na_0	na_2
na_1	na_3
na_2	ba_0
na_3	ba_1

$F^{(2)}$	$L^{(2)}$
$\$\$_0$	na_0
ba_0	na_1
ba_1	$\$\$_0$
na_0	na_2
na_1	na_3
na_2	ba_0
na_3	ba_1

$F^{(2)}$	$L^{(2)}$
$\$\$_0$	na_0
ba_0	na_1
ba_1	$\$\$_0$
na_0	na_2
na_1	na_3
na_2	ba_0
na_3	ba_1

suffix (array offset) range of <u>na</u>.<u>na</u>

Geometric-BWT

- O(|P|) LF-Mapping on FM-index = O(|P|/λ) LF-Mapping + 4-sided range counting on Geometric BWT λ times.
- Disadvantage: There are at most λ factorizations of pattern matching genuine occurrences on text.

Our Index (LMS-Factorization)

- LMS is short for LeftMost S-type (S^*), which is a term from one of the linear-time suffix array construction algorithms, SAIS [Ge Nong et al, DCC'09], and later an grammar compression scheme, GCIS [D.S.N Nunes et al, DCC'18].

b b a a b b c c a a \$

SL-types = L L S^* S S S L L L L S^*

LMS-factor

$T^{(LMS)}$ = b | a n | a n | a b | a n | a n a | \$

SL-types = L | S^* L | S^* L | S^* L | S^* L | S^* L L | S^*

Our Index (LMS-Factorization)

- There are at most 2 different factorizations of P, $P^{(\text{LMS, S})}$ and $P^{(\text{LMS, L})}$.

$$P = \text{anana}$$

$$\begin{array}{l} P_S^{(\text{LMS})} = \text{a n} \mid \text{a n} \mid \text{a} \\ \text{SL-types} = \text{L L} \mid \text{S}^* \text{ L} \mid \text{S}^* \end{array}$$

$$\begin{array}{l} P_L^{(\text{LMS})} = \text{a n} \mid \text{a n a} \\ \text{SL-types} = \text{L L} \mid \text{S}^* \text{ L L} \end{array}$$

Our Index (LMS-Factorization + λ-factorization)

- LMS-factors could be long substring of T, we further apply λ-factorization on each LMS-factor of T.
- The lex(icographical) & colex(icograpical) grammar rules, G_{lex} and G_{colex} , are mapping factors in set of factors to their lex. and colex. rank.

T = bananabanana\$
$\lambda = 2$

G_{lex}	G_{colex}
0 : \$	0 : \$
1 : a	1 : a
2 : ab	2 : b
3 : an	3 : ba
4 : b	4 : na

factor

$T^{(LMS,\,2)}$ = b | a n | a n | a b | a n | a n . a | \$
SL-types = L | S* L | S* L | S* L | S* L | S* L . L | S*

$T^{(LMS,\,2,\,Glex)}$ = 4 3 3 2 3 3 1 0

For simplicity, we take $T^{(LMS,\,2)}$ for the purpose of presentation

Our Index (LMS-Factorization + λ-factorization)

- There are at most 2 different LMS-factorizations of P, $P_S^{(LMS)}$ and $P_L^{(LMS)}$.
- There are at most λ factorizations of first LMS-factor of $P_S^{(LMS)}$ and $P_L^{(LMS)}$.

P = anana

$P_S^{(LMS,\ 1)}$ = a . n | a n | a
$P_S^{(LMS,\ 2)}$ = a n | a n | a
SL-types = L L | S^* L | S^*

$P_L^{(LMS,\ 1)}$ = a . n | a n . a
$P_L^{(LMS,\ 2)}$ = a n | a n . a
SL-types = L L | S^* L L

Note that $P_S^{(LMS,\ 1)} = P_L^{(LMS,\ 1)}$ and $P_S^{(LMS,\ 2)} = P_L^{(LMS,\ 2)}$ in terms of final factorization in this sample pattern, but it is not the case in general.

Our Index (Counting Multiple-LMS-factor Pattern)

T = bananabanana\$

P = anana

$\lambda = 2$

$T^{(\mathrm{LMS},\,2)}$ = b | a n | a n | a b | a n | a n . a | \$

$P_S^{(\mathrm{LMS},\,1)}$ = * a . n | a n | a * ---> $P_S^{(\mathrm{LMS},\,1)}$ = * a . n | a n | a *

$F^{(\mathrm{LMS},\,2)}$	$L^{(\mathrm{LMS},\,2)}$
$\$_0$	a_0
a_0	an_0
ab_0	an_1
an_0	an_2
an_1	an_3
an_2	ab_0
an_3	b_0
b_0	$\$_0$

$F^{(\mathrm{LMS},\,2)}$	$L^{(\mathrm{LMS},\,2)}$
$\$_0$	a_0
a_0	an_0
ab_0	an_1
an_0	an_2
an_1	an_3
an_2	ab_0
an_3	b_0
b_0	$\$_0$

There is no n in the range

Our Index (Counting Multiple-LMS-factor Pattern)

T = bananabanana\$
P = anana
$\lambda = 2$

$T^{(\mathrm{LMS},\,2)} = \mathrm{b} \mid \mathrm{a\,n} \mid \mathrm{a\,n} \mid \mathrm{a\,b} \mid \mathrm{a\,n} \mid \mathrm{a\,n\,.\,a} \mid \$$

$P_S^{(\mathrm{LMS},\,2)} = *\ \mathrm{a\,n} \mid \mathrm{a\,n} \mid \mathrm{a}\ *$ ---> $P_S^{(\mathrm{LMS},\,2)} = *\ \mathrm{a\,n} \mid \mathrm{a\,n} \mid \mathrm{a}\ *$

$F^{(\mathrm{LMS},\,2)}$	$L^{(\mathrm{LMS},\,2)}$
$\$_0$	a_0
a_0	an_0
ab_0	an_1
an_0	an_2
an_1	an_3
an_2	ab_0
an_3	b_0
b_0	$\$_0$

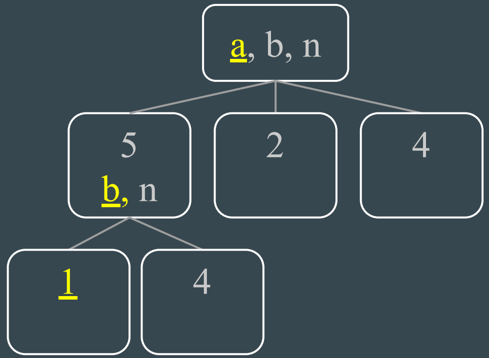
$F^{(\mathrm{LMS},\,2)}$	$L^{(\mathrm{LMS},\,2)}$
$\$_0$	a_0
a_0	an_0
ab_0	an_1
an_0	an_2
an_1	an_3
an_2	ab_0
an_3	b_0
b_0	$\$_0$

Our Index (Counting Single-LMS-factor Pattern)

- If LMS-factorization of pattern results in only single LMS-factor and λ-factorization on the LMS-factor results in multiple factors, counting is following the algorithm applied on Geometric-BWT.
- If final factorization of pattern is single factor, there is a generalized suffix trie, *GST*, built from all factors on $T^{(LMS, \lambda)}$ along with count information in nodes of the *GST* to deal with such scenario.

$T^{(LMS, 2)}$ = b | a n | a n | a b | a n | a n . a | $

P = ab



Experimental Result (Dataset)

- Cere, E.Coli, and Para from Repetitive Corpus on Pizza & Chili Corpus
- Chr19.15 is concatenation of 15 different human chromosome 19 processed from chr19.1000.fa on tudocomp dataset
- Artificial.x are artificial texts generated for simulating concatenated generational genome sequences, where the mutation rate is x%.

Experimental Result (Construction Time)

dataset	FM-index time [s]	RLFM$^{(0)}$ time [s]	RLFM$^{(1)}$ λ	RLFM$^{(1)}$ time [s]
CERE	101.6	102.6	1	881.3
			2	471.4
			3	365.0
			4	299.4
			5	290.4
			6	277.4
			7	275.9
			8	276.6
CHR19.15	207.7	208.9	1	2093.7
			2	1107.7
			3	807.3
			4	683.0
			5	633.2
			6	626.6
			7	609.6
			8	597.2
E.COLI	24.5	25.0	1	187.1
			2	110.3
			3	87.1
			4	71.9
			5	66.7
			6	65.8
			7	66.3
			8	71.1
PARA	98.0	98.0	1	755.7
			2	410.5
			3	319.3
			4	267.4
			5	260.9
			6	251.5
			7	248.2
			8	252.2

dataset	FM-index time [s]	RLFM$^{(0)}$ time [s]	RLFM$^{(1)}$ λ	RLFM$^{(1)}$ time [s]
ARTIFICIAL.1	130.0	131.4	1	616.1
			2	339.4
			3	263.2
			4	227.4
			5	224.1
			6	230.7
			7	230.5
			8	233.7
ARTIFICIAL.2	129.2	132.9	1	578.2
			2	328.3
			3	254.5
			4	218.7
			5	213.2
			6	220.5
			7	220.3
			8	224.2
ARTIFICIAL.4	130.5	137.1	1	555.2
			2	555.2
			3	243.4
			4	213.6
			5	209.7
			6	218.1
			7	216.5
			8	222.8
ARTIFICIAL.8	132.4	144.4	1	530.6
			2	306.4
			3	241.0
			4	211.0
			5	207.1
			6	214.9
			7	215.5
			8	219.7

RLFM [Mäkinen & Navarro CPM'05]

Experimental Result (Space)

input text			RLFM$^{(0)}$		RLFM$^{(1)}$				
name	space [MiB]	σ	$r^{(0)}$ [M]	space [MiB]	λ	space [MiB]	$\sigma^{(1)}$	$r^{(1)}$ [M]	lg $\|P\|$
CERE	439.9	6	11.6	26.8	1	26.5	6	11.6	-
					2	20.4	26	8.3	-
					3	18.4	95	6.7	12
					4	17.3	271	5.8	11
					5	16.7	602	5.3	11
					6	15.1	1081	5.1	12
					7	14.9	1790	5.0	13
					8	14.9	2810	4.9	13
CHR19.15	845.8	6	32.3	70.8	1	69.7	6	32.3	-
					2	54.9	22	23.5	-
					3	50.8	58	19.0	10
					4	47.1	140	16.5	9
					5	44.9	305	15.1	-
					6	40.3	620	14.3	11
					7	39.7	1174	13.8	12
					8	39.4	2086	13.6	13
E.COLI	107.5	16	15.0	26.2	1	25.4	16	15.0	-
					2	20.4	123	10.7	-
					3	19.3	399	8.5	-
					4	17.8	809	7.3	13
					5	17.2	1272	6.7	12
					6	15.3	1764	6.3	13
					7	15.1	2356	6.2	13
					8	15.1	3251	6.1	-
PARA	409.4	6	15.6	34.4	1	34.0	6	15.6	-
					2	26.5	26	11.3	-
					3	24.5	96	9.1	12
					4	22.6	296	7.9	11
					5	20.0	774	7.2	11
					6	19.6	1620	6.9	12
					7	19.4	2701	6.7	13
					8	19.3	4013	6.7	14

input text			RLFM$^{(0)}$		RLFM$^{(1)}$				
name	space [MiB]	σ	$r^{(0)}$ [M]	space [MiB]	λ	space [MiB]	$\sigma^{(1)}$	$r^{(1)}$ [M]	lg $\|P\|$
ARTIFICIAL.1	502.5	5	50.9	91.4	1	89.5	5	50.9	-
					2	84.9	17	39.0	8
					3	71.2	51	32.3	7
					4	69.2	131	28.7	8
					5	68.0	297	26.7	9
					6	67.4	611	25.8	10
					7	67.1	1164	25.3	11
					8	66.9	2058	25.1	11
ARTIFICIAL.2	500.0	5	87.5	141.8	1	138.5	5	87.5	-
					2	131.5	17	67.0	7
					3	111.8	51	55.5	7
					4	109.9	131	49.3	7
					5	108.5	297	45.9	9
					6	107.8	611	44.3	9
					7	107.4	1164	43.5	10
					8	107.2	2069	43.2	11
ARTIFICIAL.4	495.0	5	147.0	215.8	1	210.1	5	147.0	-
					2	198.7	17	111.4	6
					3	169.7	51	91.8	6
					4	168.4	131	81.1	7
					5	167.0	297	75.4	8
					6	166.0	611	72.6	9
					7	165.4	1164	71.3	10
					8	165.0	2077	70.7	11
ARTIFICIAL.8	485.0	5	237.4	300.2	1	290.9	5	237.4	-
					2	286.4	17	174.3	8
					3	239.6	51	141.3	7
					4	235.1	131	123.4	7
					5	231.0	297	113.9	8
					6	228.4	611	109.2	9
					7	226.6	1164	107.0	10
					8	225.5	2079	106.1	11

Future Work

- Is λ-factorization required?
- How to augment the current index to support locate(P)?
- Comparison of count(P) to FM-index family (e.g. RLFM+ [Sirén et al. SPIRE’08], Faster-Minuter [Gog et al. DCC’16])
- Comparison of locate(P) to GCIIS family (e.g. Akagi et al, SPIRE’21, Díaz et al. SPIRE’21)

Feedback or Question?