修士学位論文要約（令和 4 年 3 月）

パラメタ化 BW 変換のオンライン構築に関する研究

橋本 大輝
指導教員：篠原 歩

Online Construction of Parameterized Burrows-Wheeler transform

Daiki HASHIMOTO
Supervisor: Ayumi SHINOHARA

Burrows-Wheeler Transform (BWT) is a string conversion method, which is used in data compression and full text indexes. Parameterized Burrows-Wheeler Transform (pBWT) generalizes BWT for parameterized matching. Two strings are a parameterized match if there is a bijection on the symbols that makes two strings equal. We propose the first online construction algorithm for pBWT. Our algorithm works in $O(n|\Pi|\frac{\log n}{\log\log n})$ time by using Navarro's dynamic data structure.

1. はじめに

Burrows-Wheeler 変換（BW 変換）とは，Burrows と Wheeler[2] によって提案された文字列可逆変換手法の一種である．BW 変換後の文字列には，同じ文字が連続して出現しやすくなり連長圧縮が効きやすいという圧縮に関わる性質と，厳密照合問題に対する索引構造としての性質が存在しており，幅広い分野で研究されている．パラメタ化一致とは，定数文字集合を Σ，変数文字集合を Π とした時，2 つの文字列 $T, S \in (\Sigma \cup \Pi)^*$ において，$\alpha \in \Sigma$ の場合，$f(\alpha) = \alpha$ を満たし，$f(T[i]) = S[i]$ を満たす全単射の写像 $f : (\Sigma \cup \Pi) \to (\Sigma \cup \Pi)$ が存在することを指す．パラメタ化照合問題とは，テキスト T とパターン P を入力とし，P とパラメタ化一致をする T の位置を求める問題である．パラメタ化照合問題は，DNA や RNA の構造解析や，プログラムの剽窃検出など文字列の構造などに着目した問題に応用される．パラメタ化 BW 変換は，Ganguly ら [3] によって提案された，BW 変換をパラメタ化照合問題に拡張した文字列変換手法である．Ganguly らのパラメタ化 BW 変換は，パラメタ化接尾辞木を用い行うものであり，その後，Kim ら [4] によって，直前符号化の定義を変更することにより簡潔なパラメタ化 BW 変換手法が提案された．

2. パラメタ化 BW 変換

パラメタ化 BW 変換は，Ganguly ら [3] によって提案された，BW 変換をパラメタ化照合問題に拡張した文字列変換手法である．パラメタ化 BW 変換の一例を表 1 に示す．パラメタ化 BW 変換を行う際には，表 1 に示したように大きく 5 つの段階を経る．まず，変換対象とする文字列 T を巡回する．その後，各巡回した文字列に対して直前符号化を行う．この直前符号化は，Baker[1] に提案されたパラメタ化照合問題を解く際に，用いられる符号化である．直前符号化後にソートし，ソートした後の行列を $rot(T)$ と定義し，$rot(T)$ の各行を復号した行列を $\overline{rot}(T)$ と定義する．最後に，$\overline{rot}(T)$ の各行に対して種類数符号化を行い，各行の末尾をつなげた文字列を文字列 T のパラメタ化 BW 変換後の文字列 $pBWT(T)$ として定義する．また，各行の先頭をつなげた文字列を $F(T)$ と定義する．

3. 本研究の成果

本研究では，文字列 T のパラメタ化 BW 変換後の文字列 $pBWT(T)$ をオンラインで構築する初めての手法を提案する．オンライン構築とは，変換対象である文字列 T が末尾から 1 文字ずつ入力され，入力毎にパラメタ化 BW 変換後の文字列を更新することで，最終的にパラメタ化 BW 変換後の文字列 $pBWT(T)$ を構築する手法である．また，本手法は Ganguly らによって定義された手順に則って構築するものではない．本提案手法では，新たに $LCP_\infty(T)$ 配列というものを定義し，使用する．$LCP_\infty(T)$ 配列とは，$rot(T)[i]$ と $rot(T)[i+1]$ の最長共通接頭辞中に出現した ∞ の個数を i 番目の要素に格納した配列である．$LCP_\infty(T)$ 配列の一例を表 2 に示す．提案手法では，文字列 T のパラメタ化 BW 変換後の文字列 $pBWT(T)$ が構築されており，文字列 T の先頭に文字 α が追加されたとき，次に示す 3 つの段階を経ることで，文字列 $\alpha \circ T$ のパラメタ化 BW 変換後の文字列 $pBWT(\alpha \circ T)$ のオンライン構築を

表 1. 文字列 $T =$ XYZZaYYZ$ のパラメタ化 BW 変換の様子

	巡回	直前符号化		$rot(T)$	$\overline{rot}(T)$	$F(T)$	$[\![\overline{rot}(T)]\!]$	$pBWT(T)$
1	XYZZaYYZ$	∞∞∞1a414$		$∞∞∞1a414	$XYZZaYYZ	$	$3212a133	3
2	YZZaYYZ$X	∞∞1a414$∞		a∞1∞$∞441	aYYZ$XYZZ	a	a133$3212	2
3	ZZaYYZ$XY	∞1a∞14$∞4		∞$∞∞41a41	Z$XYZZaYY	3	3$3212a13	3
4	ZaYYZ$XYZ	∞a∞14$∞44		∞a∞14$∞44	ZaYYZ$XYZ	2	2a133$321	1
5	aYYZ$XYZZ	a∞1∞$∞441	→	∞1a∞14$∞4	ZZaYYZ$XY	1	12a133$32	2
6	YYZ$XYZZa	∞1∞$∞441a	整列	∞1∞$∞441a	YYZ$XYZZa	1	133$3212a	a
7	YZ$XYZZaY	∞∞$∞441a4		∞∞$∞441a4	YZ$XYZZaY	3	33$3212a1	1
8	Z$XYZZaYY	∞$∞∞41a41		∞∞1a414$∞	YZZaYYZ$X	2	212a133$3	3
9	$XYZZaYYZ	$∞∞∞1a414		∞∞∞1a414$	XYZZaYYZ$	3	3212a133$	$

表 2. 文字列 $T =$ XaYZZaZYZa$ の $LCP_\infty(T)$ 配列と $rot(T)$ の様子

	$LCP_\infty(T)$	$rot(T)$
1	0	$∞a∞∞1a252a
2	0	a$∞a∞∞1a252
3	2	a∞∞1a252a$∞
4	0	a∞∞2a$∞a661
5	1	∞a$∞a∞61a25
6	1	∞a2∞2a$∞a66
7	1	∞a∞∞1a252a$
8	1	∞1a2∞2a$∞a6
9	2	∞∞a$∞a661a2
10	2	∞∞1a252a$∞a
11	-	∞∞2a$∞a661a

行う．なお，2 つの文字列 T, S の結合を $T \circ S$ で表す．最初に，文字列 $\alpha \circ T$ の種類数符号化を行う．その後，$pBWT(\alpha \circ T), F(\alpha \circ T)$ を構築し，最後に $LCP_\infty(\alpha \circ T)$ 配列を構築する．上記の流れを，1 文字読み込むごとに行うことで，文字列全体のパラメタ化 BW 変換をオンラインで行うことができる．また，本手法では Navarro ら [5] によって提案された動的なデータ構造を用いることで，下記に示す定理 1，2，3 の計算時間で文字列 $\alpha \circ T$ のパラメタ化 BW 変換後の文字列 $pBWT(\alpha \circ T)$ のオンライン構築を行う．また，変数文字のアルファベットサイズを $|\Pi|$ とする．

定理 1 文字列 T の種類数符号化から文字列 $\alpha \circ T$ の種類数符号化の構築は，$O(|\Pi|)$ 時間で行える．

定理 2 $pBWT(T), F(T)$ から $pBWT(\alpha \circ T), F(\alpha \circ T)$ の構築は，$O(|\Pi|\frac{\log n}{\log\log n})$ 時間で行える．

定理 3 $LCP_\infty(T)$ 配列から $LCP_\infty(\alpha \circ T)$ 配列の構築は $O(|\Pi|\frac{\log n}{\log\log n})$ 時間で行える．

したがって，変数文字のアルファベットサイズを $|\Pi|$ とすると，長さ n の文字列 T のパラメタ化 BW 変換 $pBWT(T)$ を $O(n|\Pi|\frac{\log n}{\log\log n})$ 時間でオンライン構築することができる．

4. まとめと今後の課題

本研究では，パラメタ化 BW 変換後の文字列をオンラインで構築する初めての手法を提案した．本手法では，1 文字ずつの入力に対して構築を行うが，単語や特定の区切り毎の入力に対してより効率的な計算時間での構築が今後の課題である．

文献

1) Brenda S Baker. A theory of parameterized pattern matching: algorithms and applications. In *Proceedings of the 25th annual ACM symposium on Theory of computing*, pp. 71–80, 1993.
2) Michael Burrows and David Wheeler. A Block-Sorting Lossless Data Compression Algorithm. Technical report, DIGITAL SRC RESEARCH REPORT, 1994.
3) Arnab Ganguly, Rahul Shah, and Sharma V Thankachan. pBWT: Achieving succinct data structures for parameterized pattern matching and related problems. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 397–407. SIAM, 2017.
4) Sung-Hwan Kim and Hwan-Gue Cho. Simpler FM-index for parameterized string matching. *Information Processing Letters*, Vol. 165, p. 106026, 2021.
5) Gonzalo Navarro and Yakov Nekrich. Optimal dynamic sequence representations. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 865–876. SIAM, 2013.