Linking BWT and XBW via Aho-Corasick Automaton: Applications to Run-Length Encoding
Bastien Cazaux, Eric Rivals

To cite this version:

HAL Id: lirmm-02382066
https://hal-lirmm.ccsd.cnrs.fr/lirmm-02382066
Submitted on 27 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Linking BWT and XBW via Aho-Corasick Automaton: Applications to Run-Length Encoding

Bastien Cazaux
Department of Computer Science, University of Helsinki, Finland
L.I.R.M.M., CNRS, Université Montpellier, France
https://www.mv.helsinki.fi/home/cazaux/
bastien.cazaux@helsinki.fi

Eric Rivals
L.I.R.M.M., CNRS, Université Montpellier, France
https://www.lirmm.fr/~rivals/
rivals@lirmm.fr

Abstract
The boom of genomic sequencing makes compression of sets of sequences inescapable. This underlies
the need for multi-string indexing data structures that helps compressing the data. The most
prominent example of such data structures is the Burrows-Wheeler Transform (BWT), a reversible
permutation of a text that improves its compressibility. A similar data structure, the eXtended
Burrows-Wheeler Transform (XBW), is able to index a tree labelled with alphabet symbols. A link
between a multi-string BWT and the Aho-Corasick automaton has already been found and led to a
way to build a XBW from a multi-string BWT. We exhibit a stronger link between a multi-string
BWT and a XBW by using the order of the concatenation in the multi-string. This bijective link
has several applications: first, it allows one to build one data structure from the other; second, it
enables one to compute an ordering of the input strings that optimises a Run-Length measure (i.e.,
the compressibility) of the BWT or of the XBW.

2012 ACM Subject Classification Mathematics of computing → Discrete mathematics; Theory of
computation → Randomness, geometry and discrete structures; Theory of computation → Data
structures and algorithms for data management

Keywords and phrases Data Structure, Algorithm, Aho-Corasick Tree, compression, RLE

Digital Object Identifier 10.4230/LIPIcs.CPM.2019.24


Supplement Material Source code at: https://framagit.org/bcazaux/compressbwt

Funding Support from the Institut de Biologie Computationnelle (ANR-11-BINF-0002).
Eric Rivals: thanks the GEM Flagship project funded from Labex NUMEV (ANR-10-LABX-0020).

1 Introduction

A seminal, key data structure, which was used for searching a set of words in a text, is the
Aho-Corasick (AC) automaton [1]. Its states form a tree that indexes all the prefixes of the
words, and each node in the tree is equipped with another kind of arc, called a Failure Link.
A failure link of a node/prefix \( v \) points to the node representing the largest proper suffix of \( v \)
in the tree. In a way, the Aho-Corasick automaton can be viewed as a multi-string indexing
data structure.

In the early 90’s, the Burrows-Wheeler Transform (BWT) of a text \( T \), which is a reversible
permutation of \( T \), was introduced for the sake of compressing a text. Indeed, the BWT
permutation tends to group identical symbols in runs, which favours compression [5]. However,
the BWT can also be used as an index for searching in \( T \), using the Backward Search
procedure [10]. In fact, the BWT of $T$ is the last column of a matrix containing all cyclic-shifts of $T$ sorted in lexicographical order. As sorting the cyclic shifts of $T$ is equivalent to sorting its suffixes, there exists a natural link between the suffix array of $T$ and the BWT of $T$. Starting in 2005, the radical increase in textual data and in biological sequences raised the need for multi-string indexes. In the multi-string case, a simple approach is to concatenate all input strings and separate them using the same termination symbol (which does not belong to the alphabet), and then to index the result with a traditional indexing data structure (e.g., a suffix array or a FM-index). Another approach is to add a unique termination symbol for each string without concatenation [3, 21]. In both approaches, an initial order on the strings is given by the permutation of the strings in the first approach, or by the order between the termination symbols in the second. All our results are made with the first approach (which minimises the alphabet size) in mind, and can be adapted to the second. Such multi-string indexes are heavily exploited in bioinformatics: first, to index all chromosomes of a genome [17], or a large collection of similar genomes, which allows aligning sequence reads simultaneously to several reference genomes [19], or second, to store and mine whole sets of raw DNA/RNA sequence reads for the purpose of comparing biological conditions or of identifying splice junctions in RNA [7, 14]. In fact, managing compressed and searchable read data sets is now crucial for bioinformatic analyses.

Initially viewed as a simple extension of single-string BWT construction, the efficient construction of multi-string BWT is not trivial and has been investigated per se. Bauer et al. proposed, among others, a lightweight incremental algorithm for their construction [3]. Then, Holt et al. devised algorithms for directly merging several, already built multi-string BWTs efficiently [14, 15], which has recently been improved to build simultaneously the companion Longest Common Prefix (LCP) table [8], or to scale up to terabyte datasets [23].

The notion of BWT has been extended into the XBW to index trees whose arcs are labelled by alphabet symbols [9]. The XBW comprises two arrays, which compactly represent the tree and offer navigational operations.

Recently, Gagie et al. propose the notion of Wheeler graphs to subsume several variants of the BWT, including the XBW of a trie for a set of strings [12]. The relation between a multi-string BWT and the XBW representation of the Aho-Corasick automaton has already been studied and exploited. Hon et al. first use the XBW representation of the Aho-Corasick trie to speed-up dictionary matching [16] building up on [4]. Manzini gave an algorithm that computes the failure links for the trie using the Suffix Array and LCP tables, and an algorithm to build the XBW of the trie with failure links from the BWT of a multi-string [22]. However, none of these established a bijective link between a multi-string BWT and the XBW of the Aho-Corasick automaton. To generalise these results, one needs to consider the order in which the strings are concatenated to form the multi-string. This idea enables us to exhibit a bijection between a multi-string BWT and the XBW of the Aho-Corasick automaton, which allows building one structure from the other in either direction (from BWT to XBW or from XBW to BWT). Finally, we exploit this bijection to find an optimal string order that minimises a Run-Length Encoding (i.e., maximises the compressibility) of these two data structures. Numerous proofs are in appendix (A preprint version of this work is available as [6]).

2 Preliminaries

In this section, we introduce basic concepts and notation, present the Burrows-Wheeler Transform and an extension of it for a set of strings. Finally, we provide a notation for the Aho-Corasick automaton and the eXtended Burrows-Wheeler Transform.
2.1 General notation

Set and Permutation. Let $i$ and $j$ be two integers such that $i \leq j$. The interval $[i,j]$ is the set of all integers between $i$ and $j$. An integer interval partition of $[i,j]$ is a set of intervals $\{[i_1,j_1], \ldots, [i_n,j_n]\}$ such that $i_1 = i$, $j_n = j$, and for all $k \in [1,n-1]$, $j_k + 1 = k_{k+1}$. We also define the order $<$ on intervals such that for two intervals $u := [i,j]$ and $v := [i',j']$, $u < v$ if $j < i'$. Let $E$ be a finite set and let $\#E$ denote its cardinality. A permutation of $E$ is an automorphism of $E$. A permutation $\sigma$ of $E$ is said to be circular if for all $i$ and $j \in E$, there exists a positive integer $k$ such that $\sigma^k(i) = j$ (where $\sigma^1(i) = \sigma(i)$ and $\sigma^k(i) = \sigma(\sigma^{k-1}(i))$). For a circular permutation $\sigma$ of $E$ and an element $e \in E$, we denote by $\sigma_e$ the function from $[1, \#E]$ to $E$ such that for all $i \in [1, \#E]$, $\sigma_e(i) := \sigma(i+\#E-1)(e)$. Given a total order $<$ on $E$, we define $<_\sigma$ as the order on $E$ such that $e <_\sigma f$ if $\sigma(e) < \sigma(f)$ for any $e, f \in E$.

String. Let $\Sigma$ be a finite alphabet. A string $w$ of length $n$ over $\Sigma$ is a sequence of symbols $w[1] \ldots w[n]$ where $w[i] \in \Sigma$ for all $i \in [1,n]$. $\Sigma^*$ is the set of all finite strings over $\Sigma$. The length of a string $w$ is denoted by $|w|$. A substring of $w$ is written as $w[i,j] := w[i] \ldots w[j]$ for some $i \leq j$. A prefix of $w$ is a substring of $w$ starting at position $1$, and a suffix of $w$ is a substring of $w$ ending at position $|w|$. A prefix $x$ (or a suffix) of a string $y$ is said proper is $x$ is different from $y$. The reverse of a string $w$, denoted by $\overline{w}$, is the string $w[n] \ldots w[2]w[1]$. We define the lexicographic order $<$ on strings as usual.

Ordered Set of Strings. Let $S = \{s_1, \ldots, s_n\}$ be a set of strings. The norm of $S$, denoted $||S||$, is the sum of the length of strings of $S$, i.e. $||S|| := \Sigma_{s \in S}|s|$. Let Prefix($S$), (respectively, Suffix($S$)) denote the set of all prefixes (resp. all suffixes) of strings of $S$. We denote by $\hat{S}$ the set of all reverse strings of strings of $S$, i.e. $\hat{S} := \{\overline{s_1}, \ldots, \overline{s_n}\}$. An ordered set of strings $P$ is a pair $(S, \sigma)$ where $S$ is a set of strings in lexicographic order, and $\sigma$ a circular permutation of $S$. We denote by $P, S$ the set of strings $S$ and by $P, \sigma$ the circular permutation $\sigma$. We denote by $\hat{P}$ the pair $(\hat{P}, S, \sigma)$.

2.2 Burrows-Wheeler Transform

BWT of a string. Let $w$ be a string and $i$ be an integer satisfying $1 \leq i \leq |w|$. The Suffix Array (SA) of $w$ [20], denoted $SA(w)$, is the array of integers that stores the starting positions of the $w[i]$ suffixes of $w$ sorted in lexicographic order. The Burrows-Wheeler Transform (BWT) [5] of $w$, denoted BWT($w$), is the array containing a permutation of the symbols of $w$ which satisfies BWT($w$)[i] := $w[SA(w)[i] - 1]$ if $SA(w)[i] > 1$, and BWT($w$)[i] := $w[i]$ otherwise. The Longest Common Prefix table (LCP) [5] of $w$, denoted LCP($w$), is the array of integers such that LCP($w$)[i] equals the length of the longest common prefix between the suffixes of $w$ starting at positions $SA(w)[i]$ and $SA(w)[i-1]$ if $i > 0$, and 0 otherwise.

For any string $s \in \Sigma^*$ and any symbol $c \in \Sigma$, one defines the functions denoted rank and select as follows: rank$_c(s, i)$ is the number of occurrences of $c$ in $s[1,i]$, and select$_c(s, j)$ is the position of the $j^{th}$ occurrence of $c$ in $s$. The arrays BWT($w$) and LCP($w$) can be computed in $O(|w|)$ time [5, 10]. Simultaneously, one can compute rank and select for BWT($w$) at no additional cost and implement them such that any rank or select query takes constant time [13, 11, 18]. We use such state-of-the-art data structure to store a BWT.

Let $C$ denote the array of length $\#\Sigma$ such that $C[c]$ equals the number of symbols of $w$ that are alphabetically strictly smaller than $c$. The Last-To-First column mapping (LF) [10] of $w$ is the function such that for any $1 \leq i \leq |w|$ one has LF($w$)[i] := $C[BWT(w)[i]] + \text{rank}_{BWT(w)[i]}(BWT(w), i)$. It is proved that $SA(w)[LF(w)[i]] = SA(w)[i] - 1$ if $SA(w)[i] \geq 2$ and $SA(w)[LF(w)[i]] = |w|$ otherwise [5, 10].
**BWT of a set of strings.** From now on, let $P := (S, \sigma)$ be an ordered set of strings. We assume the symbol $\$ is not in $\Sigma$ and is alphabetically smaller than all other symbols. We denote by $m_P$ the string obtained by concatenating the strings of $P.S := \{s_1, \ldots, s_{|P.S|}\}$ separated by a $\$ and following the order $P.\sigma$. i.e., $m_P := s_{P.\sigma(1)} \$ s_{P.\sigma(2)} \$ \cdots \$ s_{P.\sigma(|P.S|)} \$. (See Figure 1). We extend the notion of BWT of a string to an ordered set of strings $P$: the BWT of $P$ is the BWT of the string $m_P$, i.e., $\text{BWT}(P) := \text{BWT}(m_P)$. We extend similarly the LF function by setting that $\text{LF}(P) := \text{LF}(m_P)$ (by using rank on $\text{BWT}(P)$).

![Figure 1 Running example with $P.S := \{abaa, abba, baba, bbaa\}$ and the corresponding $m_P$.](image)

We define the **Longest Representative Suffix** table ($\text{LRS}$) of $P$ as the array of $|m_P|$ integers satisfying: for any $i \in [1, |m_P|]$ one has $\text{LRS}(P)[i] := \text{select}_{\$}(m_P|\text{SA}(m_P)[i] : |m_P|, 1) - 1$. The entry $\text{LRS}(P)[i]$ gives the length of the substring of $m_P$ starting at position $\text{SA}(m_P)[i]$ up to, but not including the next $. Using the $\text{LRS}$ table, we extend the notion of LCP table to an ordered set of strings. For $P$, we set $\text{LCP}(P)[i] := \min(\text{LCP}(m_P)[i], \text{LRS}(P)[i])$ for any $i \in [1, |m_P|]$. Using tables $\text{BWT}(P)$ and $\text{LCP}(m_P)$, we can compute the tables $\text{LRS}(P)$ and $\text{LCP}(P)$ in linear time in $|m_P|$ (see Appendix). Figure 2 illustrates the LCP and LRS tables.

### 2.3 Aho-Corasick automaton and eXtended Burrows-Wheeler Transform

**Tree.** Let $T$ be a tree and $u$ be a node of $T$. Let $\bot$ denote the root of $T$. We denote by $\text{Parent}_T(u)$ the parent of $u$ in $T$, by $\text{Children}_T(u)$ the set of children of $u$ in $T$, and by $\text{Leaves}_T(u)$ the set of leaves in the subtree of $u$ in $T$. Let $v$ be a leaf of $T$; we denote by $B_T(v)$ the subtree of $T$ containing all nodes comprised between $\bot$ and $v$ included. As for a leaf $v$ in the subtree of $u$ in $T$, $\#\text{Children}_{B_T(v)}(u) = 1$, we denote by $\text{Child}_{B_T(v)}(u)$ the unique element of $\text{Children}_{B_T(v)}(u)$. Given a total order $\prec$ on $\text{Leaves}_{B_T(\bot)}$, we extend this order on the set $\text{Children}_{B_T(v)}$ for any node $v$ of $T$ by: for any $x, y$ in $\text{Children}_{B_T(v)}$, $x \prec y$ iff $\min_{y' \in \text{Leaves}_{B_T(v)}} x' \prec \min_{y' \in \text{Leaves}_{B_T(v)}} y'$.

**Aho-Corasick tree.** The **Aho-Corasick automaton** ($\text{AC}$) $[1]$ of a set of strings $S$ is a digraph whose set of nodes is the set of all prefixes of the strings of $S$. This graph is composed of two trees on the same node set. The first tree, which we called the **Aho-Corasick Tree** ($\text{ACT}$), has an arc from a prefix $u$ to a different prefix $v$ if $u$ is the longest proper prefix of $v$ in $\text{Prefix}(S)$ (see Figure 4). The second tree, termed **Aho-Corasick Failure link** ($\text{ACFL}$), has an arc from a prefix $u$ to a different prefix $v$ if $v$ is the longest proper suffix of $u$ in $\text{Prefix}(S)$.

**eXtended Burrows-Wheeler Transform.** Let $T$ be an ordered tree (i.e., with a total order on the set of children of each node) such that every node of $T$ is labelled with a symbol from an alphabet $\Sigma$. We define the functions $\delta$ and $\pi$ on the set of nodes of $T$ except for the root such that for a node $v$ of $T$, $\delta(v)$ is the label of $v$, and $\pi(v)$ is the string obtained by concatenating the labels from $v$’s parent up to the root of $T$. Let $\prec$ be the total order between the nodes of $T$ such that for $u$ and $v$ two nodes of $T$, $u \prec v$ if $\pi(u)$ is strictly lexicographically smaller than $\pi(v)$ or $u$ is before $v$ in the order of $T$. 
In [9], the eXtended Burrows-Wheeler Transform (XBWT) of a tree $T$ is an array of symbols of $\Sigma$ of length $\#PA(T)$, such that the entry at position $i$ gives the label of the node $PA(T)[i]$. The eXtended Burrows-Wheeler Last (XBWL) of a tree $T$ is the bit array of length $\#PA(T)$ such that $XBWL(T)[i]$ equals 1 if the node $PA(T)[i]$ is the last child of its parent, and 0 otherwise. Figure 5a illustrates XBWT and XBWL.

**eXtended Burrows-Wheeler Transform of Aho-Corasick automaton.** For a set of strings $S := \{s_1, \ldots, s_n\}$, we denote by $S^8$ the set $\{s_1 \$, \ldots, s_n \$$\}$. Let $P$ be an ordered set of strings. We define $ACT(P)$ as the variant of the Aho-Corasick tree of $P.S^8$ where the label of an arc is shifted on the deepest node of this arc (i.e., the label of a node $v$ of $ACT(P)$ is the label of $(u, v)$ in $ACT(P.S)$ where $u$ is the parent of $v$ in $ACT(P)$) and equipped with the order $<_{P.S}$. Indeed, $<_{P.S}$ is the order on the leaves satisfying: for $u$ and $v$ two leaves of $ACT(P)$, $u <_{P.S} v$.

\[ <_{P.S} := \text{the order on the leaves satisfying: for } u \text{ and } v \text{ two leaves of } ACT(P), u <_{P.S} v \]

1 In [9], the XBW-transform is defined as $XBW(T)[i] := (XBWT[i], XBWL[i])$, for any position $i$. 

**Example 1.** With the tree of Figure 3, we have $\delta(\text{a}) = b$ and $\pi(\text{a}) = aba = \delta(\text{a})\delta(\text{a})\delta(\text{a})$, and also $\delta(\text{b}) = b$ and $\pi(\text{b}) = baa = \delta(\text{b})\delta(\text{b})\delta(\text{a})$. Thus, $\text{a} < \text{b} < \text{a}$ since $\pi(\text{a}) <\pi(\text{b})$.

**Figure 2** Tables $LCP(P)$, $LRS(P)$, $SA(m_P)$ and $BWT(P)$ for the running example. One has $LCP(P)[10] = LRS(P)[10] = 2$ although the longest common prefix between suffixes of rank 9 and 10 is $aasab$ of length 5 (i.e., although $LCP(m_P)[10] = 5$).
We define $\pi$ which is impossible. This concludes the proof.

---

**Link between BWT and XBW of the Aho-Corasick Automaton**

Here, we introduce a decomposition of a multi-string BWT that leads us to exhibit a bijection with the Aho-Corasick automaton (see Proposition 2). This builds on and extends Manzini’s work [22]. We extend this bijection to the XBWT of the Aho-Corasick automaton (see Figure 4). For space reasons, many proofs are given in the Appendix (or the preprint [6]).

**Proposition 2.** Dec._Pre is a bijection between Decom.BWT($P$) and Prefix(Decomp.$P.S$).

Let us start the proof of Proposition 2 with the following Lemma.

**Lemma 3.** Let $u = [i, j] \in$ Decom.BWT($P$). For all $k \in [i, j]$:


**Proof.** Let us show by contraposition that $LRS(P)[k - 1] = LCP(P)[k]$ for all $k \in [i + 1, j]$. Assume that there exists $k \in [i + 1, j]$ such that $LRS(P)[k - 1] \neq LCP(P)[k]$. Whenever $LRS(P)[k - 1] < LCP(P)[k]$, we get by definition that $LRS(P)[k] \geq LCP(P)[k]$, and thus $LRS(P)[k - 1] < LRS(P)[k]$. By the definition of $LRS(P)$, $m_P[SA(m_P)[k - 1] + LRS(P)[k - 1]] = \$. By the definition of $LCP(P)$, for all $j \in [1, LCP(P)[k - 1] - 1]$, $m_P[SA(m_P)[k] + j] = m_P[SA(m_P)[k - 1] + j]$. As $LRS(P)[k - 1] < LCP(P)[k]$, we have $m_P[SA(m_P)[k] + LRS(P)[k - 1]] = m_P[SA(m_P)[k - 1] + LRS(P)[k - 1]] = \$, which is impossible since $LRS(P)[k - 1] < LRS(P)[k]$. Whenever $LRS(P)[k - 1] > LCP(P)[k]$, as $LCP(P)[k] = LRS(P)[k]$, the string $m_P[SA(m_P)[k] : m_P[n]]$ is lexicographically strictly smaller than $m_P[SA(m_P)[k - 1] : m_P[n]]$, which is impossible. This concludes the proof.
Proof. Let \( u = [i, j] \) be an interval of \( \text{Decomp}_\text{BWT}(P) \). First, we prove that \( \text{Dec}_\text{Pre}[u] \in \text{Prefix}(\overleftarrow{P.S}) \), and then to prove the bijection, we show \( \text{Dec}_\text{Pre} \) is injective and surjective.

By definition, \( \text{LRS}(P)[i] = \text{select}_4(m_P[\text{SA}(m_P)[i]] : |m_P|, 1) - 1 \), we get that \( m_P[\text{SA}(m_P)[i] + \text{LRS}(P)[i]] = \). Hence, \( m_P[\text{SA}(m_P)[i] + \text{LRS}(P)[i] - 1] = m_P[\text{SA}(m_P)[i] + \text{LRS}(P)[i] - 1] = m_P[\text{SA}(m_P)[k] : \text{SA}(m_P)[k] + \text{LRS}(P)[k] - 1] \). Hence, we have \( \text{LCP}(P)[k] = \text{LRS}(P)[i] \). Therefore by the definition of \( \text{Decomp}_\text{BWT}(P) \), we get \( u_1 = u_2 \).

Injectiveness. Let \( u_1 = [i_1, j_1] \) and \( u_2 = [i_2, j_2] \) be two elements of \( \text{Decomp}_\text{BWT}(P) \).

Without loss of generality, we take \( i_1 \leq i_2 \). Assume that \( \text{Dec}_\text{Pre}[u_1] = \text{Dec}_\text{Pre}[u_2] \), we have that \( m_P[\text{SA}(m_P)[i_1] : \text{SA}(m_P)[i_1] + \text{LRS}(P)[i_1] - 1] = m_P[\text{SA}(m_P)[i_2] : \text{SA}(m_P)[i_2] + \text{LRS}(P)[i_2] - 1] \) and thus for all \( k \in [i_1, i_2] \), \( m_P[\text{SA}(m_P)[i_1] : \text{SA}(m_P)[i_1] + \text{LRS}(P)[i_1] - 1] = m_P[\text{SA}(m_P)[k] : \text{SA}(m_P)[k] + \text{LRS}(P)[k] - 1] \). Hence, we have \( \text{LCP}(P)[k] = \text{LRS}(P)[i_1] \) and \( \text{LRS}(P)[k] = \text{LRS}(P)[i_1] \).

By Proposition 2, there exists an integer interval partition of \([1, |m_P|]\) (i.e., \( \text{Decomp}_\text{BWT}(P) \)) that is in bijection with the set of nodes of \( \text{AC}(\overleftarrow{P.S}) \). The following extends this bijection by building an isomorphic graph of \( \text{AC}(P.S) \) whose set of nodes is \( \text{Decomp}_\text{BWT}(\overleftarrow{P}) \). This states how to simulate an Aho-Corasick automaton using the BWT (similarly to [22]).

\textbf{Theorem 4.} Using tables \( \text{BWT}(\overleftarrow{P}), \text{LCP}(\overleftarrow{P}), \text{LRS}(\overleftarrow{P}) \) and the function \( \text{LF}(\overleftarrow{P}) \), we can build the graph \( (\text{Decomp}_\text{BWT}(\overleftarrow{P}), \text{A}_\text{T}(\overleftarrow{P}) \cup \text{A}_\text{F}(\overleftarrow{P}) ) \) that is isomorphic to \( \text{AC}(P.S) \).

\[ \text{A}_\text{T}(P) := \{(u, v) \in \text{Decomp}_\text{BWT}(P)^2 \mid \not\exists x \in u \text{ such that } \text{LF}(P)[x] \neq 0 \text{ and } \text{LF}(P)[x] \in v \}, \]

\[ \text{A}_\text{F}(P) := \{(u, v) \in \text{Decomp}_\text{BWT}(P)^2 \mid \max_{k < i \leq \text{LRS}(P)[v]} (k) \in v \text{ with } u := [i, j] \}. \]

More precisely, the graph \( (\text{Decomp}_\text{BWT}(\overleftarrow{P}), \text{A}_\text{T}(\overleftarrow{P})) \) is isomorphic to the tree \( \text{ACT}(P.S) \) and the graph \( (\text{Decomp}_\text{BWT}(\overleftarrow{P}), \text{A}_\text{F}(\overleftarrow{P})) \) is isomorphic to the tree \( \text{ACFL}(P.S) \) (see Appendix).

\textbf{Link between BWT and XBW.} After giving a new proof of the relation between Aho-Corasick automaton and BWT, we exhibit a new (bijective) link between the BWT and the XBW, which takes into account the order in the multi-string (Theorem 5). This leads to both, another construction algorithm of the XBW from the BWT, and to a construction of the BWT from the XBW, and thereby extends Manzini’s results (Corollary 6).

Alike the definition of \( \text{Decomp}_\text{BWT}(P) \) for an ordered set of strings \( P \), we define \( \text{Decomp}_\text{XBW}(T) \) of a tree \( T \) of \( t \) nodes as the integer interval partition of \([1, t] \) such that

\[ [i, j] \in \text{Decomp}_\text{XBW}(T) \text{ if and only if } \begin{cases} \text{XBWL}(T)[k] = 1, & \text{for } k \in \{i - 1, j\} \\ \text{XBWL}(T)[k] = 0, & \text{for } k \in \{i, j - 1\}. \end{cases} \]
Theorem 5 (See Figure 5b). There exists a bijection, denoted BWT_XBW, between Decompt_BWT(P) and Decompt_XBW(ACT(P)) such that for all \( u \in \text{Decomp}_B(P) \) with \( u := [i, j] \), BWT_XBW(u) := [i', j'] and \( z := \text{Parent}_{\text{ACT}(P)}(\text{PA}(\text{ACT}(P))[i']) \):

1. Let \( \{y_1, y_2, \ldots, y_{\#u}\} := \text{Leaves}_{\text{ACT}(P)}(z) \) such that \( y_k <_{P, \sigma} y_l \Rightarrow k < l \); then

   \[
   \text{BWT}(P)[i, j] = \delta(\text{Child}_{\text{ACT}(P)}(y_1)(z))\delta(\text{Child}_{\text{ACT}(P)}(y_2)(z))\ldots\delta(\text{Child}_{\text{ACT}(P)}(y_{\#u})(z)).
   \]

2. Let \( \{x_1, x_2, \ldots, x_{\#(BWT_XBW(u))}\} := \text{Children}_{\text{ACT}(P)}(z) \) such that \( x_k <_{P, \sigma} x_l \Rightarrow k < l \); then

   \[
   \text{XBW}(\text{ACT}(P))[[i', j']] = \delta(x_1)\delta(x_2)\ldots\delta(x_{\#(BWT_XBW(u))}).
   \]

Theorem 5 provides us with a strong link between BWT(P) and XBW(ACT(P)), which allows transforming one structure into the other. This leads to the following corollary.

Corollary 6. Let P be an ordered set of strings.

1. Using tables BWT(P), LCP(P) and LRS(P) of an ordered set of strings P, we can build the tables XBWT(ACT(P)) and XBWL(ACT(P)) in linear time of \( ||P,S|| \times \# \Sigma \).

2. Using tables XBWT(ACT(S)) and XBWL(ACT(S)) of a set of strings S, we can build the tables BWT(S), LCP(S) and LRS(S) in linear time of \( ||S|| \times \# \Sigma \) where P is an ordered set of strings such that P,S = S.

The idea behind the algorithms is to exploit the link of Theorem 5 to compute each substring of the BWT or of the XBWT associated to each element of Decompt_BWT or of Decompt_XBW.

4 Optimal ordering of strings for maximising compression

In this section, we show how to use Theorem 5 to compute the permutation that leads to a BWT and a XBWT having the minimum Run Length Encoding.
4.1 Minimum permutation problem for BWT and XBWT

Run-Length Encoding [24] is a widely used method to compress strings. For a string $w$, the Run-Length Encoding splits $w$ into the minimum number of substrings containing a single symbol. The size of the Run-Length Encoding of $w$ is the cardinality of the minimum decomposition. For example, $abaaaccabbb = a^1b^2a^3c^2a^1b^3$ (where the power notation $\alpha^n$ means $n$ copies of symbol $\alpha$), the size of the Run-Length Encoding is 6 (for the decomposition has 6 blocks).

We define Run-Length measures: $d_B$ for a BWT, and $d_X$ for a XBW (similar to those of [15]). For $P$ an ordered set of strings, let $d_B(P)$ be the cardinality of the set $\{i \in [1, \#\text{BWT}(P) - 1] \mid \text{BWT}(P)[i] \neq \text{BWT}(P)[i + 1]\}$. Similarly, let $d_X(P)$ be the cardinality of the set $\{i \in [1, \#\text{XBWT}(\text{ACT}(\mathcal{P})) - 1] \mid \text{XBWT}(\text{ACT}(\mathcal{P}))[i] \neq \text{XBWT}(\text{ACT}(\mathcal{P}))[i + 1]\}$.

Given two ordered sets of strings, $P_1$ and $P_2$ such that $P_1.S = P_2.S$ (i.e., they contain the same set of strings), Theorem 5 implies that their BWT may differ, and thus $d_B(P_1)$ and $d_B(P_2)$ may also differ. We define the following minimisation problems. As the Run-Length Encoding of BWT($P$) has size $d_B(P) + 1$, finding an optimal solution of Min-Permutation-BWT can help compressing BWT($P$).

**Definition 7** (Min-Permutation-BWT and Min-Permutation-XBWT). Let $S$ be a set of strings. The problem Min-Permutation-BWT asks for an ordered set of strings $P$ that minimises $d_B(P)$ and such that $P.S = S$. The problem Min-Permutation-XBWT asks for an ordered set of strings $P$ that minimises $d_X(P)$ and such that $P.S = S$.

To simplify Min-Permutation-BWT, we consider specific ordered sets of strings. Let $P$ be an ordered set of strings and let $\perp$ denote the root of $\text{ACT}(\mathcal{P})$. We say that $P$ is topologically planar if for each node $u$ in $\text{ACT}(\mathcal{P})$ and $v \in \text{Leaves}_{\text{ACT}(\mathcal{P})}(\perp) \setminus \text{Leaves}_{\text{ACT}(\mathcal{P})}(u)$, there does not exist $u_1$ and $u_2$ in $\text{Leaves}_{\text{ACT}(\mathcal{P})}(u)$ such that $u_1 <_{P,\sigma} v <_{P,\sigma} u_2$. In other words, $P$ is topologically planar if we can draw the tree $\text{ACT}(\mathcal{P})$ by ordering the leaves with $<_{P,\sigma}$ without arcs crossing each other.
Let $P$ be an ordered set of strings, which is not necessarily topologically planar. We denote by $P_{\sigma}$ the ordered set of strings such that $P_{\sigma}S = PS$, and such that $P_{\sigma}\sigma$ satisfies for any $u$ in $ACT(\overline{P})$, any $v \in \text{Leaves}_{ACT}(\overline{P})(u)$, and any $u_1, u_2$ in $\text{Leaves}_{ACT}(\overline{P})(u)$ such that $u_1 <_{P,\sigma} v <_{P,\sigma} u_2$, we have $u_1 <_{P,\sigma} u_2 <_{P,\sigma} v$. As we have a bijection between the set of circular permutations of $PS$ and the set of leaves of $ACT(\overline{P})$, we can unambiguously define the ordered set of strings $P_{\sigma}$ that is topologically planar.

**Proposition 8.** Let $P$ be an ordered set of strings. We have $d_B(P_{\sigma}) \leq d_B(P)$ and $d_B(P_{\sigma}) = d_X(P_{\sigma}).$

**Proof.** For the inequality, let us prove that any modification of the order used to create $P_{\sigma}$ decreases the value of $d_B$. Let $P$ be an ordered set of strings that is not topologically planar. Let $u$ in $ACT(\overline{P})$, $v \in \text{Leaves}_{ACT}(\overline{P})(u)$, and let $u_1$ and $u_2$ in $\text{Leaves}_{ACT}(\overline{P})(u)$ such that $u_1 <_{P,\sigma} v <_{P,\sigma} u_2$. Let $P'$ be the copy of $P$ where the only difference is $u_1 <_{P,\sigma} u_2 <_{P,\sigma} v$. Let $x \in \text{Decomp}\_\text{BWT}(P)$ such that $\text{Dec}\_\text{Pre}(x) = u$. By Theorem 5, for all $y \in \text{Decomp}\_\text{BWT}(P) \setminus \{x\}$, we have $\text{BWT}(P)_y[y] = \text{BWT}(P')_y[y]$, $\text{BWT}(P)_y[x] = \ldots \delta[u_1]\delta[v]\delta[u_2]\ldots$ and $\text{BWT}(P')_y[x] = \ldots \delta[u_1]\delta[v]\delta[u_2]\ldots$ with $v = \text{Child}_{ACT}(\overline{P})(u)$. As $\delta[u_1] = \delta[u_2]$ and $\delta[u_1] \neq \delta[v]$, we have $d_B(P') \leq d_B(P)$.

For the equality, it is enough to see that for any element $u$ in $\text{Decomp}\_\text{BWT}(P_{\sigma})$, the numbers of distinct successive symbols are identical in $\text{BWT}(P_{\sigma})[u]$ and in $\text{XBWT}(ACT(\overline{P}_{\sigma}))$. Thus, for two successive elements $u = [i_1, i_2]$ and $v = [i_2, i_3]$ of $\text{Decomp}\_\text{BWT}(P_{\sigma})$, we obtain an equivalence between $\text{BWT}\_\text{XBW}[u] = [i_1, j_1]$ and $\text{BWT}\_\text{XBW}[v] = [i_2, j_2]$:

$$\text{BWT}(P_{\sigma})[j_1] = \text{BWT}(P_{\sigma})[i_2] \iff \text{BWT}(ACT(\overline{P}_{\sigma}))[j_1'] = \text{BWT}(ACT(\overline{P}_{\sigma}))[i_2'].$$

Thanks to Proposition 8, we can restrict the search to ordered sets of strings that are topologically planar when solving $\text{Min-Permutation-BWT}$ or $\text{Min-Permutation-XBWT}$. Furthermore, an optimal solution of $\text{Min-Permutation-BWT}$ for $S$ is also an optimal solution of $\text{Min-Permutation-XBWT}$ for $S$, and vice versa. This yields the following theorem.

**Theorem 9.** Let $S$ be a set of strings. We can find an optimal solution for $\text{Min-Permutation-BWT}$ and for $\text{Min-Permutation-XBWT}$ in $O(|S| \times \#\Sigma)$ time.

### 4.2 Proof of Theorem 9

As a reminder, Proposition 8 states that an optimal solution of $\text{Min-Permutation-BWT}$ is also an optimal solution of $\text{Min-Permutation-XBWT}$, and vice versa. In the following of this proof, we only prove the result regarding $\text{Min-Permutation-XBWT}$.

To start, let us give an overview of the algorithm:

1. we take a random permutation $\sigma$ of $S$ and define $P$ such that $PS = S$ and $P_{\sigma} = \sigma$,
2. we build $ACT(P)$, $\text{XBWT}(ACT(P))$ and $\text{Decomp}\_\text{XBW}(ACT(P))$,
3. we find $P'$ which is an optimal solution of $\text{Min-Permutation-XBWT}$.

In the following, we define the problem $\text{Min-Permutation-Table}$ and explicit its link to $\text{Min-Permutation-XBWT}$ (Lemma 10). Lemma 10 gives us a linear time algorithm for finding an optimal solution of $\text{Min-Permutation-Table}$, and thus we can apply this algorithm to obtain an optimal solution for $\text{Min-Permutation-XBWT}$.

Given $A$ an array of symbols of $\Sigma$, we define $\text{Char}(A)$ as the set of (different) symbols in $A$. Given $T$ an array of $n$ symbols of $\Sigma$ and $D$ an integer interval partition of $[1, n]$ such for each interval $[i, j]$ of $D$, $\#(\text{Char}(T[i, j])) = j - i + 1$ (i.e., all the symbols of
Lemma 10. Let $S$ be a set of strings and let $P$ be an ordered set of strings such that $P.S = S$. For an optimal solution $T'$ of Min-Permutation-Table for $XBWT(ACT(P))$ and for $Decom_{XBW}(ACT(P))$, there exists an optimal solution $P'$ of Min-Permutation-XBWT for $S$ such that $XBWT(ACT(P')) = T'$.

Let $T$ be an array of $n$ symbols of $\Sigma$ and let $D$ be an integer interval partition of $[1, n]$ such for each interval $[i, j]$ of $D$, $\#(\text{Char}(T[i, j])) = j - i + 1$. Let $A(D)$ be the array of all intervals in $D$ in the order $<$, and $B(T, D)$ the array of size $\#(A(D)) - 1$ such that the position $i$ of $B(T, D)$ is $B(T, D)[i] = \text{Char}(A(D))[i] \cap \text{Char}(A(D))[i + 1]$. For any set of symbols $C := \{c_1, \ldots, c_m\}$ where $c_1 < \ldots < c_m$, we define $\text{word}(C)$ as the string $c_1 \ldots c_m$.

Lemma 11. Let $T$ be an array of $n$ symbols of $\Sigma$ and let $D$ be an integer interval partition of $[1, n]$ such for any interval $[i, j]$ of $D$, $\#(\text{Char}(T[i, j])) = j - i + 1$.

- If there exists $i$ in $[1, n]$ such that $[i, i] \in D$, then $T'_1[i, i - 1]T'_2$ is an optimal solution of Min-Permutation-Table for $T$ and for $D$, where $T'_1$ is an optimal solution of Min-Permutation-Table for $T[1, i]$ and for $\{|i', j'| D \mid j' \leq i\}$, and $T'_2$ is an optimal solution of Min-Permutation-Table for $T[i, n]$ and for $\{|i', j'| D \mid i' \geq i\}$.

- If there exists $i$ in $[1, \#B(T, D)]$ such that $\#(B(T, D)[i]) = 0$, then $T'_1T'_2$ is an optimal solution of Min-Permutation-Table for $T$ and for $D$, where $T'_1$ is an optimal solution of Min-Permutation-Table for $T[1, A(D)[i][1]]$ and for $\{|i', j'| D \mid j' \leq A(D)[i][1]\}$, and $T'_2$ is an optimal solution of Min-Permutation-Table for $T[A(D)[i][1] + 1][0, n]$ and for $\{|i', j'| D \mid i' \geq A(D)[i][1] + 1\}$.

- If there exists $i$ in $[1, \#B(T, D)]$ such that $\#(B(T, D)[i]) = 1$, then $T'_1aaT'_2$ is an optimal solution of Min-Permutation-Table for $T$ and for $D$, where $B(T, D)[i] = \{a\}$, $T'_1$ is an optimal solution of Min-Permutation-Table for $T[1, A(D)[i - 1][1]|\text{word(Char}(A(D))[i] \setminus \{a\})$ and for $\{|i', j'| D \mid j' < A(D)[i][1]\} \cup \{|A(D)[i][0], A(D)[i][1] - 1\}$, and $T'_2$ is an optimal solution of Min-Permutation-Table for $\text{word}(\text{Char}(A(D))[i + 1] \setminus \{a\})T[A(D)[i + 2][0, n]$ and for $\{|i', j'| D \mid i' > A(D)[i + 1][0]\} \cup \{|A(D)[i][1] + 1[0, n], A(D)[i][1]\}$.

Proof. All the proofs are derived from the equality $d_A(T[1, i], D) = d_A(T[1, i], \{|i', j'| D \mid j' \leq i\}) + d_A(T[i, n], \{|i', j'| D \mid i' \geq i\})$ for all $i \in [1, n]$.

Lemma 12. Let $T$ be an array of $n$ symbols of $\Sigma$ and let $D$ be an integer interval partition of $[1, n]$ such that for any interval $[i, j]$ of $D$, $\#(\text{Char}(T[i, j])) = j - i + 1$. Whenever for any $i \in [1, \#B(T, D)]$ one has $\#(B(T, D)[i]) \geq 2$, Algorithm 1 gives an optimal solution for Min-Permutation-Table in $O(\#T \times \#\Sigma)$ time.

Proof. The tables $A(D)$ and $B(T, D)$, we need $O(\#T \times \#\Sigma)$ time. As both tables are smaller than $T$, the for loop of Algorithm 1 also takes $O(\#T \times \#\Sigma)$ time.
Algorithm 1: Computation of an array $T'$ of symbols satisfying for any $[i, j] \in D$, $\text{Char}(T[i, j]) = \text{Char}(T'[i, j])$.

**Input**: An instance of Min-Permutation-Table $T$ and $D$

**Output**: A string $T'$

1. $\text{let} \ last \leftarrow \$ such that $\$ \notin \Sigma$;
2. $T' \leftarrow$ empty string;
3. for $i \in [1, \#B(T, D)]$ do
   1. let $\text{lettre} \leftarrow$ random element of $B(T, D)[i] \setminus \{last\}$;
   2. $T' \leftarrow T' \ \text{word}(\text{Char}(A(D)[i]) \setminus \{\text{lettre}, last\})$;
   3. $T' \leftarrow T' \ \text{lettre lettre}$;
   4. $last \leftarrow \text{lettre}$;
5. $T' \leftarrow T' \ \text{word}(\text{Char}(A(D)[\#A(D)]) \setminus \{last\})$;
6. return $T'$;

▶ **Lemma 13.** Let $T$ be an array of $n$ symbols of $\Sigma$ and let $D$ be an integer interval partition of $[1, n]$ such that for each interval $[i, j]$ of $D$, $\#(\text{Char}(T[i, j])) = j - i + 1$. The problem Min-Permutation-Table can be solved in $O(\#T \times \#\Sigma)$ time.

**Proof.** By Lemma 12 and Lemma 11, we can compute an optimal solution of Min-Permutation-Table by splitting the interval, applying Algorithm 1 on each part, and then merging the strings output by Algorithm 1.

5 Conclusion and Perspectives

Here, we present a new view of the Burrows-Wheeler Transform: as the text representation of an Aho-Corasick automaton that depends on the concatenation order. This induces a link between the Burrows-Wheeler Transform and the eXtended Burrows-Wheeler Transform, via the Aho-Corasick automaton. This link allows one to transform one structure into the other (for which we provide algorithms). We also exploit this link to find in linear time an ordering of input strings that optimises the compression of the concatenated strings.

Of course, it would be interesting to evaluate even empirically the gain of this compression on real life data. In bioinformatics, one wishes to index a collection of genomes from individual of the same species. For instance, it can be all variants of a virus genome within a host (e.g., HIV or Ebola virus in human); their number and relative frequency may change along time [2]. An application is then to compare the sequencing reads obtained from any new infected individual with this index to determine whether some variants are becoming more frequent or resistant to some treatment. Some viruses evolve rapidly to circumvent immune response, the number of potential variants can be large. Hence, it is practically relevant to reduce the index space by finding an optimal permutation of the genomes that would maximised the Run-Length Encoding of the index.

References


A Appendix

A.1 Linear time construction of tables \( LRS(P) \) and \( LCP(P) \)

To build in linear time the tables \( LRS(P) \) and \( LCP(P) \), we use Algorithm 2, which is proved in Lemma 14. We summarise this property in Proposition 15. The tables for the running example are illustrated in Figure 2.

Lemma 14. Let \( i \in [1, |m_P|] \).

\[
LRS(P)[i] = \begin{cases} 
0 & \text{if } i \in [1, \#(P.S)], \\
LRS(P)[LF(P)[i]] - 1 & \text{otherwise}.
\end{cases}
\]

Proof. For any \( i \in [1, |w|] \), we know that \( SA(w)[LF(w)[i]] = SA(w)[i] - 1 \) if \( SA(w)[i] \geq 2 \) and \( SA(w)[LF(w)[i]] = |w| \) otherwise; hence, we get

- if \( SA(w)[i] \geq 2 \),
  \[
  LRS(P)[LF(P)[i]] = \text{select}_S(m_P[SA(m_P)[LF(P)[i]] : |m_P|], 1) - 1 \\
  = \text{select}_S(m_P[SA(w)[i] - 1 : |m_P|], 1) - 1 \\
  = \begin{cases} 
0 & \text{if } m_P[SA(w)[i] - 1] = $, \\
LRS(P)[i] + 1 & \text{otherwise}.
\end{cases}
\]

- if \( SA(w)[i] = 1 \),
  \[
  LRS(P)[LF(P)[i]] = \text{select}_S(m_P[SA(m_P)[LF(P)[i]] : |m_P|], 1) - 1 \\
  = \text{select}_S(m_P[|m_P| : |m_P|], 1) - 1 \\
  = 0.
\]

We define the function \( \text{Letter} \) from \( [1, |m_P|] \) to \( \Sigma \) such that \( C[\text{Letter}[i]] < i \leq C[\text{Letter}[i + 1]] \) (see the definition of \( \text{LF} \) on page 3) where \( \Sigma = \{c_1, \ldots, c_{\#\Sigma}\} \) with \( c_1 < \ldots < c_{\#\Sigma} \). Thus, we define \( \text{RLF}(P) \) from \( [1, |m_P|] \) to \( [1, |m_P|] \) such that

\[
\text{RLF}(P)[i] = \text{select}_{\text{Letter}}^{-1}(BWT(P), i - C[\text{Letter}[i]]).
\]

Let \( i \) be an integer between 1 and \( |m_P| \). We have

\[
\text{RLF}(P)[LF(P)[i]] = \text{select}_{\text{Letter}}^{-1}(BWT(P), LF(P)[i] - C[\text{Letter}[LF(P)[i]]]) \\
= \text{select}_{\text{BWT}}^{-1}(BWT(P), LF(P)[i] - C[BWT(P)[i]]) \\
= \text{select}_{\text{BWT}}^{-1}(BWT(P), \text{rank}_{BWT}(w)[i](BWT(w), i)) \\
= i.
\]

Hence, the function \( \text{RLF}(P) \) is the reverse bijection of \( \text{LF}(P) \), and as \( m_P[SA(m_P)[i] - 1] = BWT(P)[i] \), one gets

\[
LRS(P)[i] = 0 \iff BWT(P)[\text{RLF}(P)[i]] = $ \\
\iff BWT(P)[\text{select}_{\text{Letter}}^{-1}(BWT(P), i - C[\text{Letter}[i]])] = $ \\
\iff \text{Letter}[i] = $ \\
\iff i \in [1, \#(P.S)].
\]

Therefore, we derive the desired equality, which concludes the proof.
Proposition 15. Let $P$ be an ordered set of strings. Using tables $\text{BWT}(P)$ and $\text{LCP}(m_P)$, we can compute the tables $\text{LRS}(P)$ and $\text{LCP}(P)$ in a time that is linear in $|m_P|$.

Proof. As the value of $\text{LCP}(P)$ at each position corresponds to the minimum between the values at same positions in $\text{LCP}(m_P)$ and in $\text{LRS}(P)$, we only need to prove that the table $\text{LRS}(P)$ can be computed from $\text{BWT}(P)$ in linear time.

Using Algorithm 2 and Lemma 14, we can compute table $\text{LRS}(P)$ in $O(|m_P|)$ time.

Algorithm 2: Computation of table $\text{LRS}(P)$.

\begin{algorithm}
  \textbf{Input :} The ordered set of strings $P$
  \textbf{Output :} The table $\text{LRS}(P)$
  
  Compute the table $\text{BWT}(P)$;
  Initialise $\text{LRS}$ as an empty table of length $|m_P|$;
  for $i \in [1, \#(P.S)]$ do
    position $\leftarrow i$;
    $nb \leftarrow 0$;
    $\text{LRS}[\text{position}] \leftarrow nb$;
    while $\text{BWT}(P)[\text{position}] \neq \$ do
      $\text{LRS}[\text{position}] \leftarrow nb$;
      $nb \leftarrow nb + 1$;
      position $\leftarrow \text{LF}[\text{position}]$;
  \end{algorithm}

A.2 Proof of Theorem 4

The next propositions exhibit the links between $\text{Decomp\_BWT}(P)$ and first, the arcs of $\text{ACT}(P.S)$ (Proposition 16), and second, the arcs of $\text{ACFL}(P.S)$ (Proposition 17).

Proposition 16 (See Figure 4). The graph $G_T(P) = (\text{Decomp\_BWT}(P), A_T(P))$ is isomorphic to the tree $\text{ACT}(P.S)$, where

$$A_T(P) := \{(u, v) \in (\text{Decomp\_BWT}(P))^2 \mid \exists x \in u \text{ such that } \text{LF}(P)[x] \neq 0 \text{ and } \text{LF}(P)[x] \in v\}.$$ 

Proof. First, we show that there exists a bijection between the node set of $\text{ACT}(P.S)$ and that of $G_T(P)$. We reuse the bijection $\text{Dec\_Pre}$, which served in Proposition 2. Let us show that for each arc $(u, v)$ of $A_T(P)$, $(\text{Dec\_Pre}[u], \text{Dec\_Pre}[v])$ is an arc of $\text{ACT}(P.S)$, and vice versa. Let $(u, v) \in A_T(P)$, i.e., $(u, v) \in (\text{Decomp\_BWT}(P))^2$ such that there exists $x \in u$ satisfying $\text{LF}(P)[x] \in v$.

According to [5], we know that $\text{SA}(m_p)[\text{LF}(m_p)[i]] = \text{SA}(m_p)[i] - 1$. By Lemma 14, we have $\text{LRS}(P)[\text{LF}(P)[x]] = \text{LRS}(P)[x] + 1$ for all $x$ such that $\text{LF}(P)[x] \neq 0$. With both equalities and Lemma 3, we obtain

\begin{align*}
\text{Dec\_Pre}[v] &= \text{Dec\_Pre}[u] \cdot m_p[\text{SA}(m_p)[\text{LF}(P)[x]] - 1] \\
&= m_p[\text{SA}(m_p)[\text{LF}(P)[x]] - 1] \\
&= m_p[\text{SA}(m_p)[\text{LF}(P)[x]] - 1] \\
&= m_p[\text{SA}(m_p)[\text{LF}(P)[x]] - 1] \\
\end{align*}

\begin{align*}
\text{Dec\_Pre}[u] &= \text{Dec\_Pre}[v] \cdot m_p[\text{SA}(m_p)[x] - 1 - \text{LRS}(P)[\text{LF}(P)[x]] - 1] \\
&= m_p[\text{SA}(m_p)[x] - 1 - \text{LRS}(P)[\text{LF}(P)[x]] - 1] \\
&= m_p[\text{SA}(m_p)[x] - 1 - \text{LRS}(P)[\text{LF}(P)[x]] - 1] \\
&= m_p[\text{SA}(m_p)[x] - 1 - \text{LRS}(P)[\text{LF}(P)[x]] - 1] \\
&= \text{Dec\_Pre}[v] \cdot m_p[\text{SA}(m_p)[x] - 1].
\end{align*}
The string Dec_PRE[u] is thus the longest prefix of Dec_PRE[v].

Let \((x, y)\) be an arc of \(ACT(\bar{P}, \bar{S})\). We take \(z\) a leaf in the subtree of \(ACT(\bar{P}, \bar{S})\) in \(y\). As \(x\) is the parent of \(y\) in \(ACT(\bar{P}, \bar{S})\), \(z\) is also a leaf in the subtree of \(ACT(\bar{P}, \bar{S})\) in \(x\). We take \(i \in [1, |m_P|]\) such that \(y = m_P[i : |m_P|]\). Hence \(y\) is a prefix of \(m_P[i + |z| - |y| : |m_P|]\) and \(z\) is a prefix of \(m_P[i + |z| - |x| : |m_P|]\). As \((x, y)\) is an arc of \(ACT(\bar{P}, \bar{S})\), \(|y| = |z| = 1\). Thus, choosing \(k\) such that \(SA(m_P)[k] = i + |z| - |y| + 1\), and \(u, v\) in \(Decomp_BWT(P)^2\) such that \(k \in u\) and \(LF(P)[k] \in v\), we get \(\overline{y} = Dec_PRE[u]\) and \(\overline{y} = Dec_PRE[v]\). This concludes the proof.

\[\textbf{Proposition 17 (See Figure 4).} \text{The graph } G_F(P) = (Decomp_BWT(P), A_F(P)) \text{ is isomorphic to the tree } ACFL(\bar{P}, \bar{S}), \text{ where} \]
\[
A_F(P) := \{(u, v) \in Decomp_BWT(P)^2 \mid \max_{\text{LRS}(P)[k]=\min_{k-1 \leq l \leq k}(\text{LCP}(P)[l])} (k) \in v \text{ with } u = [i, j]\}.
\]

\textbf{Proof.} First, let us show the following equivalence. Let \(u = [i, j] \in Decomp_BWT(P)\).

\[
w \in Decomp_BWT(P) \text{ such that } \exists k \in w \text{ with } k < i \text{ and } \text{LRS}(P)[k]=\min_{k-1 \leq l \leq k}(\text{LCP}(P)[l])\]
\[
\iff \text{Dec}_\text{PRE}[u] \text{ is a suffix of Dec_PRE}[u].
\]

Let \(w \in Decomp_BWT(P)\) such that \(\exists k \in w \text{ with } k < i \text{ and } \text{LRS}(P)[k]=\min_{k-1 \leq l \leq k}(\text{LCP}(P)[l])\).

Hence, we have for all \(l \in [k-1, i]\), \(\text{LRS}(P)[k] \leq \text{LCP}(P)[l]\), and thus

\[
\text{Dec}_\text{PRE}[u] = m_P[SA(m_P)[i : |m_P|] + LRS(P)[i] - 1]
\]
\[
= m_P[SA(m_P)[i : |m_P|] + \text{LRS}(P)[k] : SA(m_P)[i + \text{LRS}(P)[i] - 1]
\]
\[
= m_P[SA(m_P)[i : |m_P|] + \text{LRS}(P)[k] : SA(m_P)[i + \text{LRS}(P)[i] - 1]
\]
\[
= m_P[SA(m_P)[k : |m_P|] + LRS(P)[k] : SA(m_P)[k + LRS(P)[k] - 1]
\]
\[
= m_P[SA(m_P)[i + \text{LRS}(P)[k] : SA(m_P)[i + \text{LRS}(P)[i] - 1]
\]
\[
\text{Dec}_\text{PRE}[w].
\]

Let \(w = [i_1, j_1]\) and \(u = [i_2, j_2]\) be two elements of \(Decomp_BWT(P)\) such that \(Dec_{\text{PRE}}[w]\)

is a suffix of \(Dec_{\text{PRE}}[u]\). Hence, we have that \(m_P[SA(m_P)[i_1] : SA(m_P)[i_1 + \text{LRS}(P)[i_1] - 1]\)
is a prefix of \(m_P[SA(m_P)[i_2] : SA(m_P)[i_2] + \text{LRS}(P)[i_2] - 1]\). By the definition of \(BWT(P)\), for all \(l \in [i_1, i_2]\), \(\text{LRS}(P)[i_1] \leq \text{LCP}(P)[l]\). This concludes the proof of the equivalence. By the equivalence, given \(u\) and \(v\) in \(Decomp_BWT(P)\) such that \(Dec_{\text{PRE}}[u]\) is a suffix of \(Dec_{\text{PRE}}[v]\),

for all \(k_1 \in u\) and \(k_2 \in v\), we have \(k_1 \leq k_2\). Hence, by taking the largest \(w\) satisfying the first step of the inequality, we obtain the longest suffix and vice versa.

\[\textbf{A.3 Proof of Theorem 5} \]

\textbf{Proof.} We define \(T_B\) (resp. \(T_X\)) as the array of intervals of \(Decomp_BWT(P)\)

(resp. \(Decomp_XBW(\text{ACT}(\bar{P}))\)) sorted in the interval order. Let us prove that \(T_B\) and \(T_X\) have the same length, and that at the same position \(i\), \(T_B[i]\) and \(T_X[i]\) represent the same prefix of \(\bar{P}\). By Proposition 2, the length of \(T_B\) is \(\#\text{Prefix}(\bar{P}, \bar{S})\). By the definition of \(Decomp_XBW\), the length of \(T_X\) is the number of 1 in \(XBW(L(\text{ACT}(\bar{P})))\), i.e., the number of internal nodes of \(\text{ACT}(\bar{P})\), and thus is equal to \(\#\text{Prefix}(\bar{P}, \bar{S})\). Hence, \(T_B\) and \(T_X\) have the same cardinalities.

Let \(i \in [1, \#\text{Prefix}(\bar{P}, \bar{S})]\). By Proposition 2, \(T_B[i]\) represents the \(i\)th suffix of strings of \(P.S\) in lexicographic order. By the definition of \(XBW(L(\text{ACT}(\bar{P}))); \), all the nodes \(PA(\text{ACT}(\bar{P}))[k]\) for \(k \in T_X[i]\) have the same parent \(z\) in \(\text{ACT}(\bar{P}),\) and \(z\) represents the \(i\)th node in \(\prec \) order, i.e., the \(i\)th suffix of strings of \(P.S\) in lexicographic order.
As for a given position $i$, $T_B[i]$ and $T_X[i]$ represent the same prefix of $\bar{P}$, we define the bijection $BWT_{\text{Xbw}}$ such that for any $i$ in $[1, \#\text{Prefix}(\bar{P}, S)]$, one has $BWT_{\text{Xbw}}[T_B[i]] = T_X[i]$. 

As the tree $\text{ACT}(\bar{P})$ represents the Aho-Corasick tree of $\bar{P}.S^\dollar$, we have a bijection $b_1$ from the node set of $\text{ACT}(\bar{P})$ onto the set of prefixes of $\bar{P}.S^\dollar$. By the definition of functions $\pi$ and $\delta$ (see page 4), for any node $v$ of $\text{ACT}(\bar{P})$, we have $b_1(v) = \pi(v)\delta(v)$.

By the definition of $\text{XBwt}(\text{ACT}(\bar{P}))$, we have that for $T_X[i] = [i', j']$

$$
\#T_X[i] = \#\{ v \text{ node of } \text{ACT}(\bar{P}) \mid \pi(v) = \text{Dec}_\text{Pre}[\text{BWT}_{\text{Xbw}}^{-1}[T_X[i]]] \} = \#\{ v \text{ node of } \text{ACT}(\bar{P}) \mid v \text{ is a child of } b_1^{-1}(\text{Dec}_\text{Pre}[\text{BWT}_{\text{Xbw}}^{-1}[T_X[i]]]) \}.
$$

As $i' \in T_X[i]$ and $b_1^{-1}(\text{Dec}_\text{Pre}[\text{BWT}_{\text{Xbw}}^{-1}[T_X[i]]]) = \text{Parent}_{\text{ACT}(\bar{P})}(\text{PA}(\text{ACT}(\bar{P}))[i'])$,

$$
\#T_X[i] = \#\{ v \text{ node of } \text{ACT}(\bar{P}) \mid v \text{ is a child of } \text{Parent}_{\text{ACT}(\bar{P})}(\text{PA}(\text{ACT}(\bar{P}))[i']) \} = \#\text{Children}_{\text{ACT}(\bar{P})}(\text{Parent}_{\text{ACT}(\bar{P})}(\text{PA}(\text{ACT}(\bar{P}))[i'])).
$$

Let $\{x_1, x_2, \ldots, x_{\#T_X[i]}\}$ be the set of children of $\text{Parent}_{\text{ACT}(\bar{P})}(\text{PA}(\text{ACT}(\bar{P}))[i'])$ sorted such that $x_1 <_{P, \sigma} \ldots <_{P, \sigma} x_{\#T_X[i]}$. By the definition of $\text{XBwt}(\text{ACT}(\bar{P}))$, for any $k \in [1, \#T_X[i]]$, we have $\text{XBwt}(\text{ACT}(\bar{P}))[i' + k - 1] = \delta(x_k)$. By the definition of $\text{Bwt}(P)$, for $T_B[i] = [i'', j'']$, we get

$$
\#T_B[i] = \#\{ w \in \bar{P}.S^\dollar \mid \text{Dec}_\text{Pre}[T_B[i]] \text{ is a prefix of } w \} = \#\{ v \text{ leaf of } \text{ACT}(\bar{P}) \mid b_1^{-1}(\text{Dec}_\text{Pre}[T_B[i]]) \text{ is an ancestor of } v \text{ in } \text{ACT}(\bar{P}) \}.
$$

As $b_1^{-1}(\text{Dec}_\text{Pre}[T_B[i]]) = b_1^{-1}(\text{Dec}_\text{Pre}[\text{BWT}_{\text{Xbw}}^{-1}[T_X[i]]]) = \text{Parent}_{\text{ACT}(\bar{P})}(\text{PA}(\text{ACT}(\bar{P}))[i'])$,

$$
\#T_B[i] = \#\{ v \text{ a leaf of } \text{ACT}(\bar{P}) \mid \text{Parent}_{\text{ACT}(\bar{P})}(\text{PA}(\text{ACT}(\bar{P}))[i']) \text{ is an ancestor of } v \text{ in } \text{ACT}(\bar{P}) \} = \#\text{Leaves}_{\text{ACT}(\bar{P})}(\text{Parent}_{\text{ACT}(\bar{P})}(\text{PA}(\text{ACT}(\bar{P}))[i'])).
$$

Let $\{y_1, y_2, \ldots, y_{\#T_B[i]}\}$ be the set of leaves of the subtree of $\text{Parent}_{\text{ACT}(\bar{P})}(\text{PA}(\text{ACT}(\bar{P}))[i'])$ in $\text{ACT}(\bar{P})$ sorted such that $y_1 <_{P, \sigma} \ldots <_{P, \sigma} y_{\#T_B[i]}$. Given $k \in [i'', j'']$, we define $w_P[k]$ as the string $m_P[\text{select}_S(m_P, \text{rank}_S(m_P, \text{SA}(P)[k] - 1) + 1 : \text{SA}(P)[k] + \text{LRS}(P)[k] - 1]$. For all $k \in [i'', j'']$, the string $w_P[k]$ is a string of $P.S$. Moreover, the definition of $<_{P, \sigma}$ implies $w_P[i''] <_{P, \sigma} \ldots <_{P, \sigma} w_P[j'']$. As for $l \in [1, \#T_B[i]]$, the string $\pi(y_l)$ is also a string of $P.S$, we get $\pi(y_l) = w_P[i'' + l - 1]$. Given $x$ and $y$ in $[i'', j''] \in \text{Decomp}_B(W, P)$, we obtain the following equivalences between orders:

$$
x < y \Leftrightarrow w_P[x] <_{P, \sigma} w_P[y] \Leftrightarrow \pi(y_{x - i'' + 1}) <_{P, \sigma} \pi(y_{y - i'' + 1}) \Leftrightarrow y_{x - i'' + 1} <_{P, \sigma} y_{y - i'' + 1}.
$$

By the definition of $\text{Bwt}(P)$, for all $k \in [1, \#T_B[i]]$, we get that

$$
\text{Bwt}(P)[i' + k - 1] = m_P[\text{SA}(m_P)[i'' + k - 1] - 1] = w_P[i'' + k - 1][w_P[i'' + k - 1] - \text{LRS}[i'' + k - 1] + 1] = \pi(y_{i'' + k - 1})[w_P[i'' + k - 1] - \text{LRS}[i'' + k - 1] + 1] = \delta(\text{Child}_{\text{ACT}(\bar{P})}(y_{i'' + k - 1}))(\text{Parent}_{\text{ACT}(\bar{P})}(\text{PA}(\text{ACT}(\bar{P}))[i'])).
$$
A.4 Proof of Corollary 6

Let $P$ be an ordered set of strings. To compute tables $\text{XBWT}(\text{ACT}(\hat{P}))$ and $\text{XBWL}(\text{ACT}(\hat{P}))$ using only $\text{BWT}(P)$, $\text{LCP}(P)$ and $\text{LRS}(P)$, we first define a new table $\text{BWD}(P)$.

The Burrows-Wheeler Decomposition of $P$, denoted by $\text{BWD}(P)$, is the array of length $\#\text{Decomp}_\text{BWT}(P)$ such that for each position $i$, $\text{BWD}(P)[i]$ is the cardinality of the $i^{th}$ element of $\text{Decomp}_\text{BWT}(P)$ in interval order.

Lemma 18. Using tables $\text{LCP}(P)$ and $\text{LRS}(P)$, Algorithm 3 computes $\text{BWD}(P)$ in linear time in $\|P.S\|$ and the table $\text{BWD}(P)$ can be stored with $\|P.S\| \times \log(\#(P.S))$ bits.

Proof of Lemma 18. For each $i$ in $[1,|m_P|]$, at the beginning of the loop for, we have that $\text{LCP}(P)[i-j] \neq \text{LRS}(P)[i-j]$ and for all $k \in [i-j+1,i-1]$, $\text{LCP}(P)[k] \neq \text{LRS}(P)[k]$. Hence, if $\text{LCP}(P)[i] \neq \text{LRS}(P)[i]$, the interval $[i-j,i-1]$ is an element of $\text{Decomp}_\text{BWT}(P)$ and the cardinality of $[i-j,i-1]$ is $j$. Otherwise, we increase $j$ by 1 because the position $i$ does not correspond to a new interval of $\text{Decomp}_\text{BWT}(P)$. For the complexity, as each step of the loop can be computed in constant time, Algorithm 3 computes $\text{BWD}(P)$ in linear time in $\|P.S\|$. As for each position $i$ of $\text{BWD}(P)$, $\text{BWD}(P)[i]$ represents the number of strings of $P.S$ having as suffix $\text{Dec}_\text{Pre}[u]$, where $u$ is the $i^{th}$ element of $\text{Decomp}_\text{BWT}(P)$ sorted in interval order, it follows that $\text{BWD}(P)[i] \leq \#(P.S)$. This concludes the proof.

Algorithm 3: Computation of table $\text{BWD}(P)$.

\begin{verbatim}
Input : The tables $\text{LCP}(P)$ and $\text{LRS}(P)$
Output : The table $\text{BWD}(P)$
\text{BWD} ← empty list;
\text{j} ← 0;
for \text{i} ∈ [1,|m_P|] do
    if $\text{LCP}(P)[i] \neq \text{LRS}(P)[i]$ then
        insert \text{j} at the end of \text{BWD};
        \text{j} ← 1;
    else \text{j} ← \text{j} + 1 ;
return \text{BWD};
\end{verbatim}

Lemma 19. Using tables $\text{BWT}(P)$ and $\text{BWD}(P)$, Algorithm 4 computes the tables $\text{XBWT}(\text{ACT}(\hat{P}))$ and $\text{XBWL}(\text{ACT}(\hat{P}))$ in linear time of $\|P.S\| \times \#\Sigma$.

Proof of Lemma 19. Algorithm 4 is an application of Theorem 5.

Using Algorithm 3 to build $\text{BWD}(P)$ and Algorithm 4, we can compute tables $\text{XBWT}(\text{ACT}(\hat{P}))$ and $\text{XBWL}(\text{ACT}(\hat{P}))$ in linear time of $\|P.S\| \times \#\Sigma$.

We define the equivalent of $\text{BWD}$ for $\text{Decomp}_\text{XBW}(P)$. The $\text{eXtended Burrows Wheeler Decomposition}$ ($\text{XBWD}$) of a tree $T$ is the array of length $\#\text{XBWL}(T)$ such that for any position $i$, $\text{XBWD}(P)[i]$ equals the cardinality of the $i^{th}$ element of $\text{Decomp}_\text{XBW}(T)$ sorted in interval order.

Lemma 20. From the table $\text{XBWL}(\text{ACT}(S))$, Algorithm 5 computes $\text{XBWD}(\text{ACT}(S))$ in linear time in $\|P.S\|$ and the table $\text{XBWD}(\text{ACT}(S))$ can be stored in $\|P.S\| \times \log(\#\Sigma)$ bits.
Algorithm 4: Computation of tables $\text{XBWT}(\text{ACT}(P))$ and $\text{XBWL}(\text{ACT}(P))$.

**Input**: The tables $\text{BWT}(P)$ and $\text{BWD}(P)$

**Output**: The tables $\text{XBWT}(\text{ACT}(P))$ and $\text{XBWL}(\text{ACT}(P))$

Initialise both $\text{XBWT}$ and $\text{XBWL}$ as empty lists; $nb \leftarrow 1$;

for $i \in [1, \#\text{BWD}(P)]$ do

$D \leftarrow$ Dictionary such that for all $l \in \Sigma$, $D[l] \leftarrow true$;

$begin \leftarrow nb$;

$last \leftarrow begin + \text{BWD}(P)[i] - 1$;

$nb \leftarrow last + 1$;

for $j \in [begin, last]$ do

if $D[\text{BWT}(P)[j]]$ then

insert $\text{BWT}(P)[j]$ at the end of $\text{XBWT}$;

insert 0 at the end of $\text{XBWL}$;

$D[\text{BWT}(P)[j]] \leftarrow false$;

end if

end for

$\text{XBWL}[\#\text{XBWL}] \leftarrow 1$;

return $\text{XBWT}$ and $\text{XBWL}$;

Proof. The proof of Lemma 20 is similar to that of Lemma 18. As for each position $i$ of $\text{XBWD}(\text{ACT}(S))$, $\text{XBWD}(\text{ACT}(S))[i]$ represents the number of the right extensions of the strings $\text{PA}(\text{ACT}(S))[i]$ in $S$ (i.e., the number of different strings $\text{PA}(\text{ACT}(S))[i]$ a which are substrings of a string of $S$ for some $a \in \Sigma$), we have $\text{XBWD}(\text{ACT}(S))[i] \leq \#\Sigma$. $\blacktriangle$

Algorithm 5: Computation of table $\text{XBWD}(\text{ACT}(S))$.

**Input**: The table $\text{XBWL}(\text{ACT}(S))$; **Output**: The table $\text{XBWD}(\text{ACT}(S))$

$\text{XBWD} \leftarrow$ empty list;

$j \leftarrow 1$;

for $i \in [1, \#\text{XBWL}(\text{ACT}(S))]$ do

if $\text{XBWL}(\text{ACT}(S)) = 1$ then

insert $j$ at the end of $\text{XBWD}$;

$j \leftarrow 1$;

else $j \leftarrow j + 1$;

end if

end for

return $\text{XBWD}$;

Lemma 21. Using tables $\text{XBWT}(\text{ACT}(S))$ and $\text{XBWD}(\text{ACT}(S))$, we can build the tables $\text{BWT}(P)$, $\text{LCP}(P)$ and $\text{LRS}(P)$ in linear time of $\|P.S\| \times \#\Sigma$, where $P$ is a topologically planar, ordered set of strings such that $P.S = S$.

Proof. In [9], Ferragina et al. prove that with both tables $\text{XBWT}(\text{ACT}(S))$ and $\text{XBWD}(\text{ACT}(S))$ one can access in constant time the children and the parents in $\text{ACT}(S)$. Hence, we can compute in linear time in $\|P.S\|$, the table $\text{TL}(\text{ACT}(S))$, where in each position of $i$ we store the number of leaves in the subtree of the node $\text{PA}(\text{ACT}(S))[i]$. We finish the proof using the results of Theorem 5 and an algorithm similar to Algorithm 4. $\blacktriangle$
A.5 Proof of Lemma 10

Proof. Let $T'$ be an optimal solution of Min-Permutation-Table for XBWT($ACT(P)$) and for Decomp_XBW($ACT(P)$). By Theorem 5, for each $[i,j] \in$ Decomp_XBW($ACT(P)$), the order of the symbols in XBWT($ACT(P)$)$[i,j]$ depends on the order on the children of the parent of PA($ACT(P)$)$[i]$. Hence, the choice of $T'$ corresponds to the choice of an order for each internal node of $ACT(P)$ over all its children. As we can extend this order to a total order on the leaves of $ACT(P)$, we can build $P'$ the ordered set of strings satisfying $P'.S = S^\$ and $P'.\sigma(\pi(f_i)) = s_i$, where the order on the leaves of $ACT(P)$ is $f_1 < \ldots < f_{\#S}$ and $s_1 < \ldots < s_{\#S}$ are the strings of $S$ in lexicographic order.