On Extensions of Maximal Repeats in Compressed Strings

Julian Pape-Lange
Technische Universität Chemnitz, Straße der Nationen 62, 09111 Chemnitz, Germany
julian.pape-lange@informatik.tu-chemnitz.de

Abstract

This paper provides upper bounds for several subsets of maximal repeats and maximal pairs in compressed strings and also presents a formerly unknown relationship between maximal pairs and the run-length Burrows-Wheeler transform.

This relationship is used to obtain a different proof for the Burrows-Wheeler conjecture which has recently been proven by Kempa and Kociumaka in "Resolution of the Burrows-Wheeler Transform Conjecture".

More formally, this paper proves that the run-length Burrows-Wheeler transform of a string S with z_S LZ77-factors has at most $73(\log_2 |S|)(z_S + 2)^2$ runs, and if S does not contain q-th powers, the number of arcs in the compacted directed acyclic word graph of S is bounded from above by $18q(1 + \log_q |S|)(z_S + 2)^2$.

2012 ACM Subject Classification Mathematics of computing → Combinatorics on words; Mathematics of computing → Combinatoric problems

Keywords and phrases Maximal repeats, Extensions of maximal repeats, Combinatorics on compressed strings, LZ77, Burrows-Wheeler transform, Burrows-Wheeler transform conjecture, Compact suffix automata, CDAWGs

Digital Object Identifier 10.4230/LIPIcs.CPM.2020.27

Related Version The preprint of this paper can be found at `https://arxiv.org/abs/2002.06265`.

Acknowledgements Fabio Cunial suggested that my previous work might be extendable from counting maximal repeats to counting extensions of maximal repeats. He also pointed out that such a result would be more interesting since it is more closely linked to the size of the compacted directed acyclic word graph. Nicola Prezza noted that my previous work also resulted in a non-trivial upper bound for the number of runs in the run-length Burrows-Wheeler transform and that a more careful investigation of the extensions of maximal repeats might result in a better bound for the Burrows-Wheeler conjecture which was unsolved at that time. I also thank Djamal Belazzougui for notifying me of the "Resolution of the Burrows-Wheeler Conjecture" by Kempa and Kociumaka.

1 Introduction

A maximal repeat P of a string is a substring such that there are two occurrences of P in the string which are preceded by different characters and succeeded by different characters. Such a pair of occurrences is called a maximal pair.

Raffinot proves in [10] that there is a natural bijection from the internal nodes in a Compacted Directed Acyclic Word Graph (CDAWG) to the maximal repeats, which is given by the labels of the paths. Also, Furuya et al. present in [7] a relation between maximal repeats and the grammar compression algorithm RePair, and they use this relation to design MR-RePair, an improved variant of RePair.

Sometimes, maximal repeats are not sufficient, since they do not contain any information about the surrounding string. Therefore, in [1], Belazzougui et al. introduce the number of right extensions of maximal repeats as a measure for the repetitiveness of strings. They further

31st Annual Symposium on Combinatorial Pattern Matching (CPM 2020).
Editors: Inge Li Gørtz and Oren Weimann; Article No. 27; pp. 27:1–27:13
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

prove that the number of arcs in the CDAWG is equal to the number of right extensions of maximal repeats and that the number of runs in the run-length Burrows-Wheeler transform (RLBWT) is bounded from above by the number of right extensions of maximal repeats.

In earlier work, I proved in [9] that the number of maximal repeats in a string S with z_S (self-referential) LZ77-factors and without q-th powers is bounded from above by $3q(z_S+1)^3-2$ and that this upper bound is tight up to a constant factor. This result implies that for a string S over an alphabet Σ, the number of arcs in the CDAWG, and thereby the number r_S of runs in the RLBWT, is bounded from above by $3|\Sigma|q(z_S+1)^3$.

We should expect that of all the $\mathcal{O}\left(q(z_S)^3\right)$ maximal repeats some provide less information than others. For example in the string

$$ba^{10}ba^{20}b\$ = baaaaaaaaaabaaaaaaaaaaaaaaaaaaaab\$,$$

we can derive all maximal pairs of the maximal repeats of a^i from the maximal pairs of a^9 and a^{19}. In this way, highly-periodic maximal repeats with exponent close to the exponent of the corresponding runs are more important than other maximal repeats which are powers of the same base.

Blumer et al. have already shown in 1987 in [2] that the CDAWG cannot compress high powers and that the CDAWG of $a^n\$$ has size $\Theta(n)$. Contrary to the CDAWG, the RLBWT does not suffer from high powers and we should expect that there are many right extensions of maximal repeats which do not increase the number of runs. And in fact, if the string is very structured, we expect that the output consists of few runs of single characters. For example Christodoulakis et al. show in [4] that the Burrows-Wheeler transform of the n-th Fibonacci string F_n is given by $b^{f_{n-2}}a^{f_{n-1}}$.

Yet, until recently, it remained an open question whether there is an upper bound for the number of runs in the RLBWT which is polynomial in the number of LZ77-factors and the logarithm of the length of the string only. This Burrows-Wheeler transform conjecture was resolved in October 2019 by Kempa and Kociumaka who prove in the first version of their arXiv-article [8] that $r_S \in \mathcal{O}\left(z_S(\log n)^2\right)$ holds and promised that they will show $r_S \in \mathcal{O}\left(\delta_S \log \delta_S \max\left(1, \log \frac{n}{\delta_S \log \delta_S}\right)\right)$ for a complexity measure $\delta_S \leq z_S$ in an extended version. In April 2020 they uploaded the extended second version to their arXiv-article [8]. In this extended version they do not only prove this tighter upper bound $r_S \in \mathcal{O}\left(\delta_S \log \delta_S \max\left(1, \log \frac{n}{\delta_S \log \delta_S}\right)\right)$, but they also prove that this upper bound is asymptotically tight for all values of n and δ_S.

This paper provides a different approach to the Burrows-Wheeler transform conjecture and shows by using maximal repeats and their extensions that $r_S \leq 73(\log_2 |S|)(z_S+2)^2$ holds.

On the way, this paper also shows that the number of arcs in the CDAWG is bounded from above by $18q(1+\log_q |S|)(z_S+2)^2$ and gives new insights into the combinatorial properties of extensions of maximal repeats which are either non-highly-periodic or cannot be extended by more than a period length.

2 Definitions

Let Σ be an *alphabet.* A *string* with *length* denoted by $|S|$ is the concatenation of *characters* $S[1]S[2]\cdots S[|S|]$ of Σ. Since it will be useful to have a predecessor and a successor for every character of the string, we also define $S[0] = \$$ and $S[|S|+1] = \$$ with $\$ \notin \Sigma$. The *substring*

$S[i..j]$ with $0 \leq i \leq j \leq |S|+1$ is the concatenation $S[i]S[i+1]\cdots S[j]$. For $i > j$ the substring $S[i..j]$ is defined to be the empty string with length 0. A *prefix* is a substring of the form $S[1..j]$ and a *suffix* is a substring of the form $S[i..|S|]$.

In this paper, we are not only interested in the substrings themselves but we are also interested in their relationship to the underlying string. We therefore use *positioned substrings*. Formally, a positioned substring is a pair (l, r) of indices and the content of the positioned substring is the substring $S[l..r]$. In order to use positioned substrings as substrings, we slightly abuse the notation in this paper and denote the positioned substrings like normal substrings with $S[l..r]$. Therefore, the term "positioned" only indicates that we are not allowed to forget the underlying indices.

An *occurrence* of a substring P is a positioned substring $S[l..r]$ such that $S[l..r] = P$ holds for the underlying substrings.

For example in the string $S = ababab$, the positioned substrings $S[1..3] = aba$ and $S[2..4] = bab$ overlap on the positioned substring $S[2..3] = ba$, but the positioned substrings $S[1..3] = aba$ and $S[4..6] = bab$ don't have a non-empty overlap. Also, in this string S, the substring $P = S[2..4] = bab$ has exactly two occurrences given by the positioned substrings $S[2..4] = bab = P$ and $S[4..6] = bab = P$.

The string S is *lexicographically strictly smaller/larger* than the string S' if S is lexicographically smaller/larger than S' and there is a mismatch $S[m] \neq S'[m]$.

A *maximal pair* of S is a triple $(n, m, l) \in \mathbb{N}^3$ with $l \geq 1$ such that $S[n..n+l-1]$ is equal to $S[m..m+l-1]$ and this property can not be extended to any side. More formally:

- $\forall i \in \mathbb{N}$ with $0 \leq i < l : S[n+i] = S[m+i]$ but
- $S[n-1] \neq S[m-1]$ and
- $S[n+l] \neq S[m+l]$.

With this notation, the string $S[n..n+l-1] = S[m..m+l-1]$ is the *corresponding maximal repeat*.

Since for a maximal pair (n, m, l) the inequality $S[n-1] \neq S[m-1]$ holds, the indices n and m cannot be equal. Also, by construction, $S[n..n+l-1]$ and $S[m..m+l-1]$ are contained in S and $S[n..n+l]$ and $S[m..m+l]$ are contained in $S\$$.

For a positioned maximal repeat $S[n..n+l-1]$, the *right-extension* of this maximal repeat is the substring $S[n..n+l]$ which is obtained by extending the maximal repeat by its successor. Similarly, the *double-sided extension* is $S[n-1..n+l]$.

Since maximal pairs are easier to handle than maximal repeats and their extensions, this paper introduces the notion of "substantially different maximal pairs" which allows to give an upper bound for the number of double-sided extensions:

Two maximal pairs (n, m, l) and (n', m', l') are *copies of each other* if the two strings $S[n-1..n+l]$ and $S[m-1..m+l]$ are equal to the two strings $S[n'-1..n'+l']$ and $S[m'-1..m'+l']$. In particular, the two maximal pairs (n, m, l) and (m, n, l) are always copies of each other. However, it is not sufficient for two maximal pairs to have identical corresponding maximal repeats in order to be copies of each other.

If two maximal pairs are not copies of each other, they are *substantially different*.

For each of the substantially different maximal pairs there can be at most two double-sided extensions of the corresponding maximal repeat. Therefore, the number of double-sided repeats is at most twice the number of substantially different maximal pairs.

A string S which is not of the form P^q for an integer $q \in \mathbb{N}_{\geq 2}$ is *primitive*, and a square S^2 with a primitive root S is a *primitively rooted square*.

A *period* of a string S is an integer p such that all characters in S with distance p are equal. A string with period length p is called *p-periodic*.

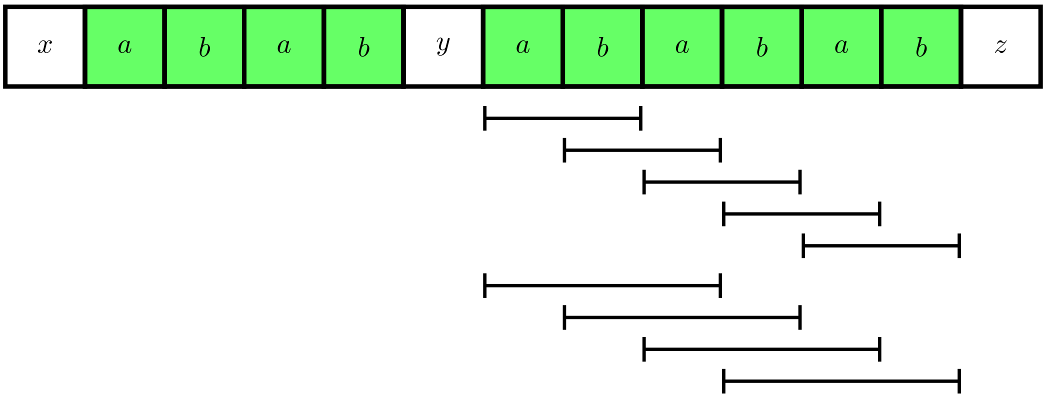

Figure 1 The string $S = xababyabababz$ with two maximal periodic extensions of the substring ab and nine extendable maximal substrings, all of them with root ab. The maximal periodic extensions are the two green substrings and each extendable substring is represented by a line indicating the underlying positions.

A string S is $\frac{1}{\geq q}$-*highly-periodic*, if it has a period with length $\frac{1}{q}|S|$ or smaller. A maximal pair is $\frac{1}{\geq q}$-*highly-periodic* if the corresponding maximal repeat is $\frac{1}{\geq q}$-highly-periodic.

For example, the strings $aaaa = a^4$, $aaaaa = a^5$ and $ababababa = (ab)^4a = (ab)^{4.5}$ are $\frac{1}{\geq 4}$-highly-periodic, but $aaaac = a^4c$ and $abababa = (ab)^{3.5}$ are not $\frac{1}{\geq 4}$-highly-periodic.

Let $S[l..r]$ be a positioned p-periodic substring with $|S[l..r]| \geq p$. The *maximal p-periodic extension* of this occurrence is the positioned substring $S[l', r']$ such that

- $l' \leq l \leq r \leq r'$,
- $S[l'..r']$ is p-periodic,
- $S[l' - 1..r']$ is not p-periodic and
- $S[l'..r' + 1]$ is not p-periodic.

With this notation, the pair $S[l' - 1, r' + 1]$ is the *padded maximal p-periodic extension*.

If p is the minimal period length of $S[l..r]$, we will omit the p and simply write *maximal periodic extension* and *padded maximal periodic extension*.

Similar to maximal pairs, two padded maximal periodic extensions $S[l - 1, r + 1]$ and $S[l' - 1, r' + 1]$ are *copies of each other* if the corresponding strings are equal. If the two padded maximal periodic extensions are not copies of each other, they are *substantially different*.

A positioned substring $S[l..r]$ with minimal period length p is *extendable* if the maximal p-periodic extension is at least $p + 1$ characters longer than $S[l..r]$. A maximal pair is *extendable*, if both occurrences of the corresponding maximal repeat are extendable.

For example, in Figure 1, we have the string $S = xababyabababz$. The positioned substrings $S[2..3] = ab$, $S[3..4] = ba$ and $S[8..11] = baba$, each with minimal period length 2, are not extendable, since their maximal periodic extensions $S[2..5] = abab$ (for both $S[2..3]$ and $S[3..4]$) and $S[7..12] = ababab$ (for $S[8..11]$) are only p characters longer. The positioned substring $S[8..8] = b$ has minimal period length 1 and therefore its maximal periodic extension is $S[8..8]$. On the other hand, the positioned substring $S[9..10] = ab$ with minimal period length 2 has the maximal periodic extension $S[7..12] = ababab$ which is 4 characters longer. Hence, the positioned substring $S[9..10]$ is extendable.

Checking all substrings, one can see that the extendable substrings of S are exactly the 9 2-periodic positioned substrings of the positioned substring $S[7..12]$ with length less than 4.

Also, the maximal pair $(7, 11, 2)$ is extendable even though both maximal periodic extensions are the same positioned substring. Also, this is the only extendable maximal pair of this string.

The (self-referential) *LZ77-decomposition* of a string S is a factorization $S = F_1F_2 \dots F_{z_S}$ in LZ77-factors, such that for all $i \in \{1, 2, \dots, z_S\}$ either

- F_i is a character which does not occur in $F_1F_2 \dots F_{i-1}$ or
- F_i is a the longest possible prefix of $S[|F_1F_2 \dots F_{i-1}| + 1..|S|]$ which occurs at least twice in $F_1F_2 \dots F_i$.

Let $\pi_i \in \{0, 1, 2, \dots, |S|\}$ be given by the lexicographic order of the cyclic permutations $S[\pi_i + 1..|S| + 1]S[1..\pi_i]$ of $S\$$. The *Burrows-Wheeler transform* defined in [3] is given by the last characters of those strings, and, since $S[0] = \$ = S[|S| + 1]$ hold by definition, these characters are given by $S[\pi_i]$.

3 Non-Highly-Periodic Maximal Pairs

The main goal of this section is to prove that in a string S the number of substantially different non-$\frac{1}{\geq 6}$-highly-periodic maximal pairs is bounded from above by $41(\log_2 |S|)(z_S + 1)(z_S + 2)$.

Along the way, this section will also prove that if S does not contain q-th powers, its CDAWG has at most $18q(1 + \log_q |S|)(z_S + 2)^2$ arcs.

In Theorem 8 of [9], I counted the number of maximal pairs around the boundaries of LZ77-factors which neither begin nor end with a power of a given exponent:

▶ **Theorem 1** (Theorem 8 of [9]). *Let S be a string. Let $F_1F_2 \dots F_zF_{z+1} = S\$$ be the LZ77-decomposition of $S\$$. Let $s_1, s_2, \dots, s_z, s_{z+1}$ be the starting indices of the LZ77-factors in $S\$$. Let $q \in \mathbb{N}_{\geq 2}$ and $i, j \in \{1, 2, \dots, z, z + 1\}$ be natural numbers.*
Then the number of different maximal pairs (n_k, m_k, l_k) such that for all k

- *the substring $S[n_k..s_i - 1]$ is not a fractional power with exponent greater than or equal to q,*
- *the substring $S[s_i..n_k + l_k - 1]$ is not a fractional power with exponent greater than or equal to q,*
- *the starting index s_i is contained in the interval $[n_k, n_k + l_k]$,*
- *the starting index s_{i+1} is not contained in the interval $[n_k, n_k + l_k]$ and*
- *the starting index s_j is contained in the interval $[m_k, m_k + l_k]$*

is bounded from above by $18q \cdot \lceil \log_q(|F_1F_2 \dots F_i|) \rceil$.

This can be slightly simplified by ignoring the underlying LZ77-structure which is not used in the proof:

▶ **Corollary 2.** *Let S be a string. Let $q \in \mathbb{N}_{\geq 2}$ be a natural number and i, j be indices of two characters in $S\$$.*

Then there are at most $18q \cdot \lceil \log_q(|S\$|) \rceil$ different maximal pairs (n_k, m_k, l_k) such that for all k

- *neither the substring $S[n_k..i - 1]$ nor the substring $S[i..n_k + l_k - 1]$ is $\frac{1}{\geq q}$-highly-periodic and*
- *the indices i and j are contained in the intervals $[n_k, n_k + l_k]$ and $[m_k, m_k + l_k]$, respectively.*

Following the proof of Theorem 8 in [9], the substring $S[i..n_k + l_k - 1]$ naturally splits into $S[n_k..i - 1]$ and $S[i..n_k + l_k - 1]$ and we can even require that the longer part(s) is/are not a high power(s). In order to have a unique longer part, we define the string $S[n_k..i - 1]$ to be longer than $S[i..n_k + l_k - 1]$, if both of these substrings have the same length.

▶ **Lemma 3.** *Let S be a string. Let $q \in \mathbb{N}_{\geq 2}$ be a natural number and i, j be indices of two characters in $S\$$.*

Then there are at most $18q(1 + \log_q |S|)$ different maximal pairs (n_k, m_k, l_k) such that for all k

- *the longer string of the substrings $S[n_k..i-1]$ and $S[i..n_k+l_k-1]$ is not $\frac{1}{\geq q}$-highly-periodic and*
- *the indices i and j are contained in the intervals $[n_k, n_k+l_k]$ and $[m_k, m_k+l_k]$, respectively.*

As proven in Lemma 4 of [9], each maximal pair has a copy such that both double-sided extensions of the corresponding maximal repeats cross LZ77-boundaries. Also, each maximal pair introduces at most two new right extensions of maximal repeats. Therefore, we can deduce a bound similar to Theorem 1 of [9] for the right extensions of maximal repeats and the arcs of the CDAWG:

▶ **Theorem 4.** *Let S be a string. Let q be further a natural number such that S does not contain q-th powers.*

Then the number of right extensions of maximal repeats in S is bounded from above by $18q(1+\log_q |S|)(z_S+2)^2-(z_S+1)$. Also, the CDAWG of S has at most $18q(1+\log_q |S|)(z_S+2)^2$ arcs.

Proof. Summing up over the first indices $i \leq j$ of the $z_S + 1$ LZ77-factors of $S\$$ yields that there are at most

$$\sum_{i=1}^{z_S+1} \sum_{j=i}^{z_S+1} 18q(1+\log_q |S|) = 9q(1+\log_q |S|)(z_S+1)(z_S+2) \leq 9q(1+\log_q |S|)(z_S+2)^2-(z_S+1)$$

substantially different maximal pairs. And since each new substantially different maximal pair introduces at most two new right extensions of maximal repeats, there are at most $18q(1 + \log_q |S|)(z_S + 2)^2 - (z_S + 1)$ different right extensions of maximal repeats.

Since the number of right extensions of (non-empty) maximal repeats is equal to the number of arcs in the CDAWG which start at internal nodes and since there are exactly $|\Sigma \cup \{\$\}| \leq z_S + 1$ arcs starting at the root, there are at most $18q(1 + \log_q |S|)(z_S + 2)^2$ arcs in the CDAWG. ◀

Additionally, there might be maximal pairs, in which the longer part(s) is/are high power(s) but the corresponding periodicity does not extend to the whole maximal repeat. In order to find a good upper bound for those maximal pairs, we need an additional lemma to limit the number of possible period lengths of prefixes and suffixes with high powers.

▶ **Lemma 5.** *Let S be a string. Let further P_1, P_2 be two substrings of S such that*

- *P_1 and P_2 are both either prefixes or suffixes of S,*
- *the length of P_2 fulfills the inequality $|P_1| \leq |P_2| \leq 2|P_1|$ and*
- *both P_1 and P_2 are $\frac{1}{\geq 3}$-highly-periodic.*

Then P_1 and P_2 have the same minimal period length.

Proof. Without loss of generality assume that P_1 and P_2 are both prefixes of S. Let p_1 and p_2 be the minimal period lengths of P_1 and P_2, respectively.

Since the inequalities $p_1 \leq \frac{1}{3}|P_1|$ and $p_2 \leq \frac{1}{3}|P_2| \leq \frac{2}{3}|P_1|$ hold, the periodicity lemma from [6] of Fine and Wilf proves, that P_1 is $\gcd(p_1, p_2)$-periodic. Since p_1 is the minimal period length of P_1, this implies that p_2 is a multiple of p_1.

However, since $P_1 \subset P_2$ and $p_2 \leq \frac{2}{3}|P_1|$ hold, a p_2-periodic base of P_2 is also p_1-periodic. Therefore $p_1 = p_2$ holds. ◀

▶ **Theorem 6.** *Let S be a string. Let i, j be indices of two characters in $S\$$.*

Then there are at most $12 \log_2 |S|$ different maximal pairs (n_k, m_k, l_k) such that for all k

- *the longer string of the substrings $S[n_k..i-1]$ and $S[i..n_k+l_k-1]$ is $\frac{1}{\geq 3}$-highly-periodic with period length p, but*
- *the substring $S[n_k..n_k+l_k-1]$ is not p-periodic and*
- *the indices i and j are contained in the intervals $[n_k, n_k+l_k]$ and $[m_k, m_k+l_k]$, respectively.*

Proof. By contradiction:

It is sufficient to prove that there are at most $6 \log_2 |S|$ different maximal pairs with the restrictions given by the prerequisites which fulfill $|S[n_k..i-1]| \geq |S[i..n_k+l_k-1]|$. By symmetry, the maximal pairs which fulfill the inequality $|S[n_k..i-1]| < |S[i..n_k+l_k-1]|$ can be bounded with an identical argument.

Assume there are at least $\lfloor 6 \log_2(|S|) \rfloor + 1$ different maximal pairs (n_k, m_k, l_k) with $|S[n_k..i-1]| \geq |S[i..n_k+l_k-1]|$ and the restrictions given by the prerequisites.

Since for all maximal pairs $1 \leq n_k$ holds, the inequality $i - n_k \leq |S\$| - 1$ holds as well. On the other hand, since $S[n_k..i-1]$ is $\frac{1}{\geq 3}$-highly-periodic, this substring has to contain at least three characters. Therefore, the inequality $3 \leq i - n_k$ holds.

Taking the logarithm yields

$$1 < \log_2(3) \leq \log_2(i - n_k) \leq \log_2(|S\$| - 1) \leq \lceil \log_2(|S|) \rceil.$$

For each maximal pair, the number $\log_2(i - n_k)$ lies in at least one of the $\lceil \log_2(|S|) \rceil - 1$ intervals $[h, h+1]$ with $1 \leq h < \lceil \log_2(|S|) \rceil$.

Using $\lceil \log_2(|S|) \rceil - 1 \leq \lfloor \log_2(|S|) \rfloor$, the pigeonhole principle now yields that there has to be a natural number L' such that

$$\left\lceil \frac{\lfloor 6 \log_2(|S|) \rfloor + 1}{\lfloor \log_2(|S|) \rfloor} \right\rceil = 7$$

of these maximal pairs have a starting index with $L' \leq \log_2(i - n_k) \leq 1 + L'$.

Therefore, for $L = 2^{L'}$ this gives a natural number L such that $L \leq i - n_k \leq 2L$ holds for each of these 7 maximal pairs.

Since the index i is contained in the interval $[n_k, n_k+l_k]$ and $|S[n_k..i-1]| \geq |S[i..n_k+l_k-1]|$ holds, the index i is also contained in the interval $[n_k + \frac{l_k}{2}, n_k + l_k]$. Hence, the inequalities $n_k + \frac{l_k}{2} \leq i$ and thereby $\frac{l_k}{2} \leq i - n_k \leq 2L$ hold. Therefore, the length l_k is at most $4L$.

Since the index j is contained in the interval $[m_k, m_k+l_k]$, this implies that the inequality $m_k \geq j - l_k \geq j - 4L$ holds. On the other hand $m_k \leq j$ so the m_k are in an interval of length $4L$.

Using the pigeonhole principle again, there are

$$\left\lceil \frac{7}{6} \right\rceil = 2$$

of these maximal pairs (n_a, m_a, l_a), (n_b, m_b, l_b) such that the distance $|m_a - m_b|$ is at most $\frac{2}{3}L$.

According to Lemma 5, both $S[n_a..i-1]$ and $S[n_b..i-1]$ have the same minimal period length. Hence, the corresponding maximal repeats are of the form $p_a P^3 s_a r_a$ and $p_b P^3 s_b r_b$ where $p_a P^3$ and $p_b P^3$ are the $|P|$-periodic parts left of i, the substrings s_a and s_b are the maximal $|P|$-periodic extensions of $p_a P^3$ and $p_b P^3$ to the right and r_a and r_b are the remaining characters of the maximal repeats.

Since the two $|P|$-periodic strings $p_a P^3 s_a$ and $p_b P^3 s_b$ starting at n_a and n_b overlap at least by $3|P|$ and since s_a and s_b are the maximal $|P|$-periodic extensions of $p_a P^3$ and $p_b P^3$, respectively, this implies that $s_a = s_b$. Therefore, the maximal repeats are of the form $p_a P^3 s r_a$ and $p_b P^3 s r_b$.

Since $|m_a - m_b| \leq \frac{2}{3}L$ holds, we can show that the $|P|$-periodic strings p_aP^3s and p_bP^3s starting at the indices m_a and m_b have at least an overlap of length $|P|$:

The strings p_aP^3s and p_bP^3s have at least the length $3|P|$. Therefore, if $P \geq \frac{L}{3}$ holds, the overlap is at least $3|P| - \frac{2}{3}L \geq |P|$.

The strings p_aP^3s and p_bP^3s also have at least the length L. Therefore, if $P \leq \frac{L}{3}$ holds, the overlap is at least $L - \frac{2}{3}L = \frac{L}{3} \geq |P|$.

In either case, the overlap is at least as long as P.

Therefore, the union of the occurrences of p_aP^3s and p_bP^3s starting at m_a and m_b is $|P|$-periodic. This implies that these two occurrences end with the same character.

If the lengths of p_aP^3s and p_bP^3s are different, this implies that both occurrences of the smaller string starting at the indices n_a and m_a or at the indices n_b and m_b are preceded by the same character which is given by the $|P|$-periodic extension to the left. This, however, implies that either (n_a, m_a, l_a) or (n_b, m_b, l_b) is not a maximal pair.

If, on the other hand, the lengths of p_aP^3s and p_bP^3s are equal, the starting indices n_a and n_b are equal and the starting indices m_a and m_b are equal as well. This, however, is only possible if either (n_a, m_a, l_a) or (n_b, m_b, l_b) is not a maximal pair or if both maximal pairs are identical.

Since both cases contradict the assumption, the assumption is wrong and the theorem is therefore true. ◀

▶ **Corollary 7.** *Let S be a string. Let $q \in \mathbb{N}_{\geq 3}$ be a natural number and i, j be indices of two characters in $S\$$.*

Then there are at most $12\left(1 + 3\frac{q}{\log_2 q}\right)\log_2|S|$ different maximal pairs (n_k, m_k, l_k) such that for all k

- *the corresponding maximal repeat $S[n_k..n_k + l_k - 1]$ is not $\frac{1}{\geq 2q}$-highly-periodic and*
- *the indices i and j are contained in the intervals $[n_k, n_k+l_k]$ and $[m_k, m_k+l_k]$, respectively.*

Also, there are at most $12\left(1 + 3\frac{q}{\log_2 q}\right)(\log_2|S|)(z_S + 1)(z_S + 2)$ different double-sided extensions of non-$\frac{1}{\geq 2q}$-highly-periodic maximal repeats.

Proof. Without loss of generality, the inequality $q \leq |S|$ holds.

If a maximal repeat $S[n_k..n_k + l_k]$ is not $\frac{1}{\geq 2q}$-highly-periodic, then either the longer of the parts $S[n_k..i - 1]$ and $S[i..n_k + l_k - 1]$ is

- not $\frac{1}{\geq q}$-highly-periodic or
- $\frac{1}{\geq q}$-highly-periodic but the corresponding periodicity does not extend to the whole maximal repeat $S[n_k..n_k + l_k]$.

Therefore, the number of different maximal pairs which fulfill the prerequisites can be bound by Lemma 3 and Theorem 6 and there are at most

$$18q(1 + \log_q|S|) + 12(\log_2|S|) \leq 36q(\log_q|S|) + 12(\log_2|S|) = 12\left(1 + 3\frac{q}{\log_2 q}\right)\log_2|S|$$

of those maximal pairs.

Summing up over the first indices $i \leq j$ of the $z_S + 1$ LZ77-factors of $S\$$ yields that there are at most

$$\sum_{i=1}^{z_S+1}\sum_{j=i}^{z_S+1} 12\left(1 + 3\frac{q}{\log_2 q}\right)\log_2|S| = 6\left(1 + 3\frac{q}{\log_2 q}\right)\log_2|S|(z_S + 1)(z_S + 2)$$

substantially different non-$\frac{1}{\geq 2q}$-highly-periodic maximal pairs.

Hence, there are at most $12\left(1 + 3\frac{q}{\log_2 q}\right)(\log_2|S|)(z_S + 1)(z_S + 2)$ different double-sided extensions of maximal repeats that are not $\frac{1}{\geq 2q}$-highly-periodic. ◀

For $q = 3$ this proves that the number of substantially different non-$\frac{1}{\geq 6}$-highly-periodic maximal pairs is bounded from above by $41(\log_2 |S|)(z_S + 1)(z_S + 2)$.

4 Highly-Periodic Maximal Pairs

The goal of this section is to prove that in a string S the number of substantially different non-extendable $\frac{1}{\geq 4}$-highly-periodic maximal pairs bounded from above by $32(\log_2 |S|)(z_S + 1)^2$.

Both occurrences of those maximal pairs, including the corresponding maximal repeat as well as the preceding and succeeding characters, are inside of the two padded maximal periodic extensions of the corresponding positioned maximal repeats.

Therefore, we will first count the number of substantially different padded maximal periodic extensions of fourth powers and the number of substantially different padded maximal periodic extensions of a given fourth power. Afterwards, we will show that each pair of padded maximal periodic extensions gives rise to at most 4 substantially different non-extendable $\frac{1}{\geq 4}$-highly-periodic maximal pairs.

We will need the "Three Squares Lemma" of Crochemore and Rytter presented in [5].

▶ **Lemma 8.** *Let u, v and w be primitive and let u^2, v^2 and w^2 be prefixes/suffixes of S with $|u| < |v| < |w|$.*

Then $|w| > |u| + |v|$ holds.

▶ **Lemma 9.** *Let S be a string and i be an index of a character in $S\$$.*

Then there are at most $4 \lfloor \log_2 |S| \rfloor$ substantially different padded maximal periodic extensions $S[l - 1..r + 1]$ of fourth powers such that $l - 1 < i \leq r + 1$.

Proof. In this proof we will only count the number of padded maximal periodic extensions $S[l - 1..r + 1]$ of fourth powers such that at least half of the interval $[l, r]$ is smaller than i, i.e. $l + \frac{r-l+1}{2} \leq i$. The other case $l + \frac{r-l+1}{2} \geq i$ is symmetrical.

Since $S[l..r]$ is at least a fourth power, the string $S[l..i - 1]$ is at least a square. Therefore, two maximal periodic extensions $S[l, r]$ and $S[l', r']$ of fourth powers have an overlap of least twice the smaller minimal period length. Therefore, if their minimal period lengths are equal, the padded maximal periodic extensions $S[l - 1, r + 1]$ and $S[l' - 1, r' + 1]$ are copies of each other. Conversely, if $S[l - 1, r + 1]$ and $S[l' - 1, r' + 1]$ are substantially different, then they have different minimal period lengths as well.

This implies that the number of substantially different padded maximal periodic extensions $S[l - 1, r + 1]$ of fourth powers such that at least half of the interval $[l, r]$ is smaller than i is bounded from above by the number of different primitively rooted squares that are suffixes of $S[1..i - 1]$.

The three squares lemma implies that for three primitively rooted squares which are suffixes of each other, the largest square is more than twice as long as the smallest square.

Since the smallest square has at least two characters and the largest square has at most $|S|$ characters, there are at most $2 \lfloor \log_2 |S| \rfloor$ primitively rooted squares which are suffixes of $S[1..i - 1]$.

Therefore, there are at most $2 \lfloor \log_2 |S| \rfloor$ padded maximal periodic extensions $S[l-1, r+1]$ of fourth powers such that at least half of the interval $[l, r]$ is smaller than i, i.e. $l + \frac{r-l+1}{2} \leq i$.

This implies that the number of padded maximal periodic extensions of fourth powers $S[l - 1, r + 1]$ such that $l - 1 < i \leq r + 1$ is bounded from above by $4 \lfloor \log_2 |S| \rfloor$. ◀

The proof also allows another useful conclusion.

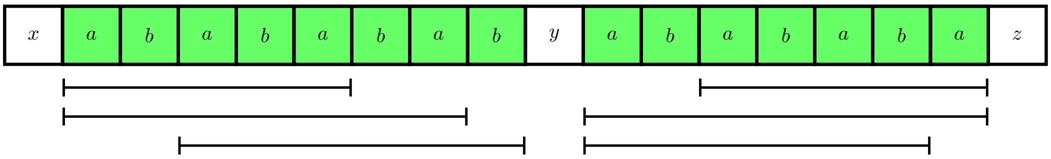

Figure 2 The string $S = xababababyabababaz$ with two maximal periodic extensions of the substring ab and the three non-extendable maximal pairs with the root ab. The maximal periodic extensions are the two green substrings and each maximal pair is represented by the two occurrences of its maximal repeat.

▶ **Corollary 10.** *Let S be a string and i be an index of a character in $S\$$. Furthermore, let P be a $\frac{1}{\geq 4}$-highly-periodic substring of S.*

Then there are at most 2 substantially different padded maximal periodic extensions $S[l-1, r+1]$ of cyclic permutations of P such that $l - 1 < i \leq r + 1$.

Combining the previous corollary with the lemma before gives rise to an upper bound of the pairs of corresponding maximal periodic extensions.

▶ **Lemma 11.** *Each pair of padded maximal periodic extensions of fourth powers which are up to cyclic rotation identical gives rise to at most 4 substantially different non-extendable $\frac{1}{\geq 4}$-highly-periodic maximal pairs.*

Proof. Each maximal pair has to be a prefix of the one padded maximal periodic extension and a suffix of the other padded maximal periodic extension, otherwise both corresponding positioned maximal repeats would be preceded or succeeded by the same character. There are two choices of which padded maximal periodic extension the corresponding positioned maximal repeat is a prefix.

For a fixed choice, the length of the maximal repeat is fixed, up to a multiple of the period length. Therefore there are only two possible lengths such that at least one of the positioned maximal repeat is not extendable.

Figure 2 shows a string with two maximal periodic extension of the substring ab and the 3 different non-extendable maximal pairs which arise from these extensions. ◀

Multiplying these upper bounds leads to the wanted upper bound:

▶ **Corollary 12.** *Let S be a string.*

Then there are at most $8(\log_2 |S|)(z_S + 1)^2$ substantially different pairs of padded maximal periodic extensions of fourth powers which are up to cyclic rotation identical.

Also, there are at most $32(\log_2 |S|)(z_S + 1)^2$ substantially different non-extendable $\frac{1}{\geq 4}$-highly-periodic maximal pairs.

5 RLBWT and Maximal Pairs

The goal of this section is to prove that the runs of the RLBWT of a string S correspond to a subset of the maximal pairs, whose size can be bound from above by $73(\log_2 |S|)(z_S + 2)^2$.

Since we are interested in the number of runs, it is useful to observe the indices i where a new run starts. These are exactly the index 1 and the indices i with $S[\pi_{i-1}] \neq S[\pi_i]$.

Since \$ occurs exactly once in $S\$$, the strings $S[\pi_{i-1} + 1..|S| + 1]$ and $S[\pi_i + 1..|S| + 1]$ have a mismatch. Also, since $S[\pi_{i-1} + 1..|S| + 1]S[1..\pi_{i-1}]$ is lexicographically smaller than $S[\pi_i + 1..|S| + 1]S[1..\pi_i]$, the string $S[\pi_{i-1} + 1..|S| + 1]$ is lexicographically strictly smaller than $S[\pi_i + 1..|S| + 1]$.

Let m be the index of the first mismatch of these two strings. With this notation, the strings $S[\pi_{i-1}+1..\pi_{i-1}+m-1]$ and $S[\pi_i+1..\pi_i+m-1]$ are equal and their predecessors as well as their successors are different. Therefore, if $m>0$, they form a maximal pair. If $m=0$, then $S[\pi_{i-1}+1]<S[\pi_i+1]$. This, however can only occur $|\Sigma|$ times.

On the other hand, since $S[\pi_{i-1}+1..\pi_{i-1}+m]$ is smaller than $S[\pi_i+1..|S|+1]S[1..\pi_i]$ and $S[\pi_i+1..\pi_i+m]$ is larger than $S[\pi_{i-1}+1..|S|+1]S[1..\pi_{i-1}]$, this maximal pair can only correspond to this pair (π_{i-1},π_i) of lexicographically neighbored cyclic permutations and the maximal pairs corresponding to different pairs (π_{j-1},π_j) of lexicographically neighbored cyclic permutations are substantially different.

▶ Remark 13. Belazzougui et al. show in Theorem 1 of [1] that the number of runs in the Burrows-Wheeler transform is even bounded in the number of right extensions of the maximal repeats. However, maximal pairs are easier to handle then right extensions of maximal repeats and we only lose a factor Σ in the worst-case by not using the right extensions.

However, while the number of maximal repeats and thereby the number of nodes in the CDAWG can be $\Theta(qz^3)$ for a suitable set of strings, the Burrows-Wheeler transform does not suffer from high powers as the CDAWG does:

▶ **Lemma 14.** *Let S be a string and let i be an index at which a new run in the Burrows-Wheeler transform starts.*

Then the maximal pair corresponding to the pair (π_{i-1},π_i) of lexicographically neighbored cyclic permutations is not extendable.

Proof. Since a maximal pair is not extendable if at least one of its corresponding positioned maximal repeats is not extendable, we have to prove that at least one of the positioned maximal repeats $S[\pi_{i-1}+1..\pi_{i-1}+m-1]$ and $S[\pi_i+1..\pi_i+m-1]$ is not extendable. Let p be the minimal period length of this maximal repeat.

Assume that the maximal p-periodic extensions of both occurrences $S[\pi_{i-1}+1..\pi_{i-1}+m-1]$ and $S[\pi_i+1..\pi_i+m-1]$ contain at least $p+1$ additional characters. In this proof, we will show that under this assumption that there is a cyclic permutation $S[w+1..|S|+1]S[1..w]$ of $S\$$ which is lexicographically between $S[\pi_{i-1}+1..|S|+1]S[1..\pi_{i-1}]$ and $S[\pi_i+1..|S|+1]S[1..\pi_i]$.

If the maximal p-periodic extension of $S[\pi_{i-1}+1..\pi_{i-1}+m-1]$ extends this occurrence to the left, the equation $S[\pi_{i-1}]=S[\pi_{i-1}+p]$ and thereby

$$S[\pi_i]\neq S[\pi_{i-1}]=S[\pi_{i-1}+p]=S[\pi_i+p]$$

holds. Therefore, the maximal p-periodic extension of $S[\pi_i+1..\pi_i+m-1]$ does not extends this string to the left. This implies that at most one of the two maximal p-periodic extensions of the occurrences $S[\pi_{i-1}+1..\pi_{i-1}+m-1]$ and $S[\pi_i+1..\pi_i+m-1]$ does extend the occurrence to the left.

Similarly, at most one of those occurrences is extended to the right by the maximal p-periodic extension.

Since, by assumption, both occurrences are p-periodically extendable, exactly one occurrence has to be p-periodically extendable to the left and exactly one occurrence has to be p-periodically extendable to the right. By symmetry we can assume without loss of generality that $S[\pi_{i-1}+1..\pi_{i-1}+m-1]$ is p-periodically extendable to the left and that $S[\pi_i+1..\pi_i+m-1]$ is p-periodically extendable to the right.

Hence, $S[\pi_{i-1}-p..\pi_{i-1}+m-1]$ and $S[\pi_i+1..\pi_i+m+p]$ are p-periodic. Also, by definition of the Burrows-Wheeler transform, the inequality $S[\pi_{i-1}+m]<S[\pi_i+m]$ holds.

Combining the periodicity with this inequality yields

$$S[\pi_{i-1}+1..\pi_{i-1}+m-1]=S[\pi_{i-1}+1-p..\pi_{i-1}+m-1-p]$$

and

$$S[\pi_{i-1}+m] < S[\pi_i+m] = S[\pi_i+m-p] = S[\pi_{i-1}+m-p]$$

which imply

$$S[\pi_{i-1}+1..|S|+1]S[1..\pi_{i-1}] < S[\pi_{i-1}+1-p..|S|+1]S[1..\pi_{i-1}-p].$$

Similarly, we get

$$S[\pi_{i-1}+1-p..\pi_{i-1}+m-1] = S[\pi_i+1..\pi_i+m-1+p]$$

and

$$S[\pi_{i-1}+m] < S[\pi_i+m] = S[\pi_i+m+p]$$

which imply

$$S[\pi_{i-1}+1-p..|S|+1]S[1..\pi_{i-1}-p] < S[\pi_i+1..|S|+1]S[1..\pi_i].$$

Since $S[\pi_{i-1}+1-p..|S|+1]S[1..\pi_{i-1}-p]$ is lexicographically between the cyclic permutations $S[\pi_{i-1}+1..|S|+1]S[1..\pi_{i-1}]$ and $S[\pi_i+1..|S|+1]S[1..\pi_i]$, these two strings are not neighbors with regard to the Burrows-Wheeler transform. This contradicts the assumption and thereby concludes the proof. ◀

Therefore, the positioned maximal repeats of the associated maximal pairs corresponding to the RLBWT are either not highly-periodic or, if they are highly-periodic, the period cannot be extended by more than a period length. This implies the following corollary and thereby leads to another proof of the Burrows-Wheeler conjecture:

▶ **Corollary 15.** *Let S be a string.*
Then, there are at most $73(\log_2|S|)(z_S+2)^2$ runs in the RLBWT.

Proof. We count the number of index-pairs (π_{i-1},π_i) where a new run starts.
Either (π_{i-1},π_i) corresponds to

- the empty maximal pair (there are at most $|\Sigma|$ of such (π_{i-1},π_i)),
- a non-$\frac{1}{\geq 6}$-highly-periodic maximal pair (there are at most $41(\log_2|S|)(z_S+1)(z_S+2)$ of such (π_{i-1},π_i)) or
- a $\frac{1}{\geq 4}$-highly-periodic non-extendable maximal pair (there are at most $32(\log_2|S|)(z_S+1)^2$ of such (π_{i-1},π_i)).

We also have to count one additional run for $i=0$.
Summing up shows that there are at most $73(\log_2|S|)(z_S+2)^2$ runs in the RLBWT. ◀

6 Conclusion

This paper proved that of the potentially $\mathcal{O}(qz^3)$ substantially different maximal pairs in a string, it is sufficient to understand a subset containing at most $73(\log_2|S|)(z_S+2)^2$ maximal pairs.

It seems therefore likely that it is possible to merge the nodes of the CDAWG which correspond maximal repeats of the same base and get a new data structure which is almost as universal and intuitive as the CDAWG but only contains $\mathcal{O}((\log|S|)(z_S)^2)$ arcs.

Also, the proofs presented in this paper do not use the underlying structure of the string. If the substrings of S and the reversed string S_{rev} are also highly compressible and have less than z' LZ77-factors each, it should be possible to prove that the number of runs in the RLBWT is bounded from above by $\mathcal{O}(z'(z_S)^2)$.

Thereby, it might be possible to derive an upper bound for the runs in the RLBWT which is only dependent on the number of LZ77-factors. Since the strings used to prove the asymptotic tightness for the upper bound $r_S \in \mathcal{O}\left(\delta_S \log \delta_S \max\left(1, \log \frac{n}{\delta_S \log \delta_S}\right)\right)$ in [8] have $z_S \in \Omega\left(\delta_S \log_2 \frac{n}{\delta_S}\right)$ LZ77-factors, such a result does not violate the asymptotic tightness.

References

1 Djamal Belazzougui, Fabio Cunial, Travis Gagie, Nicola Prezza, and Mathieu Raffinot. Composite repetition-aware data structures. In Ferdinando Cicalese, Ely Porat, and Ugo Vaccaro, editors, *Combinatorial Pattern Matching - 26th Annual Symposium, CPM 2015, Ischia Island, Italy, June 29 - July 1, 2015, Proceedings*, volume 9133 of *Lecture Notes in Computer Science*, pages 26–39. Springer, 2015. doi:10.1007/978-3-319-19929-0_3.

2 Anselm Blumer, J. Blumer, David Haussler, Ross M. McConnell, and Andrzej Ehrenfeucht. Complete inverted files for efficient text retrieval and analysis. *J. ACM*, 34(3):578–595, 1987. doi:10.1145/28869.28873.

3 M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. Technical report, DEC Systems Research Center, 1994.

4 Manolis Christodoulakis, Costas S. Iliopoulos, and Yoan José Pinzón Ardila. Simple algorithm for sorting the fibonacci string rotations. In Jiří Wiedermann, Gerard Tel, Jaroslav Pokorný, Mária Bieliková, and Július Štuller, editors, *SOFSEM 2006: Theory and Practice of Computer Science*, pages 218–225, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

5 M. Crochemore and W. Rytter. Squares, cubes, and time-space efficient string searching. *Algorithmica*, 13(5):405–425, May 1995. doi:10.1007/BF01190846.

6 N. J. Fine and H. S. Wilf. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 16(1):109–114, 1965. URL: http://www.jstor.org/stable/2034009.

7 I. Furuya, T. Takagi, Y. Nakashima, S. Inenaga, H. Bannai, and T. Kida. Mr-repair: Grammar compression based on maximal repeats. In *2019 Data Compression Conference (DCC)*, pages 508–517, March 2019. doi:10.1109/DCC.2019.00059.

8 Dominik Kempa and Tomasz Kociumaka. Resolution of the burrows-wheeler transform conjecture. *CoRR*, abs/1910.10631, 2019. arXiv:1910.10631.

9 Julian Pape-Lange. On Maximal Repeats in Compressed Strings. In Nadia Pisanti and Solon P. Pissis, editors, *30th Annual Symposium on Combinatorial Pattern Matching (CPM 2019)*, volume 128 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 18:1–18:13, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi:10.4230/LIPIcs.CPM.2019.18.

10 Mathieu Raffinot. On maximal repeats in strings. *Inf. Process. Lett.*, 80(3):165–169, 2001. doi:10.1016/S0020-0190(01)00152-1.