Burrows-Wheeler Transform on Purely Morphic Words

A. Frosini*, I. Mancini†, S. Rinaldi†, G. Romana‡ and M. Sciortino‡

*Università di Firenze
Firenze, Italy
andrea.frosini@unifi.it

†Università di Siena
Siena, Italy
ilaria.mancini@student.unisi.it
simone.rinaldi@unisi.it

‡Università di Palermo
Palermo, Italy
giuseppe.romana01@unipa.it
marinella.sciortino@unipa.it

The study of the compressibility of repetitive sequences is an issue that is attracting great interest. We consider *purely morphic words*, which are highly repetitive sequences generated by iterating a morphism φ that admits a fixed point (denoted by $\varphi^\infty(a)$) starting from a given character a belonging to the finite alphabet A, i.e. $\varphi^\infty(a) = \lim_{i\to\infty} \varphi^i(a)$. Such morphisms are called *prolongable* on a. Here we focus on the compressibility via the Burrows-Wheeler Transform (BWT) of infinite families of finite sequences generated by morphisms. In particular, denoted by $r(w)$ the number of equal-letter runs of a word w, we provide new upper bounds on $r(\texttt{bwt}(\varphi^i(a)))$, i.e. the number of equal-letter runs produced when BWT is applied on $\varphi^i(a)$. Such bounds depend on the factor complexity $f_x(n)$ of the infinite word $x = \varphi^\infty(a)$, that counts, for each $n \geq 0$, the number of distinct factors of x having length n.

More in detail, given the infinite word $x = \varphi^\infty(a)$ over the finite alphabet A, the following upper bounds for $r(\texttt{bwt}(\varphi^i(a)))$ can be proved:

1. if $f_x(n)$ is $\Theta(n)$ then $r(\texttt{bwt}(\varphi^i(a))) \in \mathcal{O}(i)$.
2. if $f_x(n)$ is $\Theta(n \log\log n)$ then $r(\texttt{bwt}(\varphi^i(a))) \in \mathcal{O}(i \log i \log\log i)$.
3. if $f_x(n)$ is $\Theta(n \log n)$ then $r(\texttt{bwt}(\varphi^i(a))) \in \mathcal{O}(i^2 \log i)$.

When the special case of binary alphabet is considered, it is possible to give an upper bound also for another class of morphisms. In particular, if φ is a morphism on the binary alphabet $\{a, b\}$ admitting the fixed point $x = \varphi^\infty(a)$ such that $f_x(n)$ is $\Theta(n^2)$, we can prove that $r(\texttt{bwt}(\varphi^i(a)) \in \mathcal{O}(i)$.

Such results allow us to state that, for the *BWT-clustering ratio* ρ [1] of almost all the sequences on the binary alphabet $\{a, b\}$ generated by a prolongable morphism φ, the following inequality holds:

$$\rho(\varphi^i(a)) = \frac{r(\texttt{bwt}(\varphi^i(a)))}{r(\varphi^i(a))} \ll 1.$$

This extends some results shown in [2].

[1] S. Mantaci, A. Restivo, G. Rosone, M. Sciortino, and L. Versari, "Measuring the clustering effect of BWT via RLE," *Theoret. Comput. Sci.*, vol. 698, pp. 79 – 87, 2017.

[2] S. Brlek, A. Frosini, I. Mancini, E. Pergola, and S. Rinaldi, "Burrows-Wheeler Transform of Words Defined by Morphisms," in *IWOCA*. 2019, vol. 11638 of *Lect. Notes Comput. Sci.*, pp. 393–404, Springer.

2022 Data Compression Conference (DCC) | 978-1-6654-7893-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/DCC52660.2022.00063

2375-0359/22/$31.00 ©2022 IEEE
DOI 10.1109/DCC52660.2022.00063

Authorized licensed use limited to: Universita degli Studi di Firenze. Downloaded on November 03,2022 at 08:35:18 UTC from IEEE Xplore. Restrictions apply.