Novel Results on the Number of Runs of the Burrows-Wheeler-Transform

Sara Giuliani[1], Shunsuke Inenaga[2], Zsuzsanna Lipták[1], Nicola Prezza[3], Marinella Sciortino[4], and Anna Toffanello[1]

[1] Dipartimento di Informatica, University of Verona, Italy,
{sara.giuliani_01, zsuzsanna.liptak}@univr.it,
anna.toffanello@studenti.univr.it

[2] Department of Informatics, Kyushu University, Fukuoka, Japan,
PRESTO, Japan Science and Technology Agency, Kawaguchi, Japan
inenaga@inf.kyushu-u.ac.jp

[3] Department of Business and Management, LUISS University, Rome, Italy,
nprezza@luiss.it

[4] Dipartimento di Matematica e Informatica, University of Palermo, Italy,
marinella.sciortino@unipa.it

Abstract. The Burrows-Wheeler-Transform (BWT), a reversible string transformation, is one of the fundamental components of many current data structures in string processing. It is central in data compression, as well as in efficient query algorithms for sequence data, such as webpages, genomic and other biological sequences, or indeed any textual data. The BWT lends itself well to compression because its number of equal-letter-runs (usually referred to as r) is often considerably lower than that of the original string; in particular, it is well suited for strings with many repeated factors. In fact, much attention has been paid to the r parameter as measure of repetitiveness, especially to evaluate the performances in terms of both space and time of compressed indexing data structures.

In this paper, we investigate $\rho(v)$, the ratio of r and of the number of runs of the BWT of the reverse of v. Kempa and Kociumaka [FOCS 2020] gave the first non-trivial upper bound as $\rho(v) = O(\log^2(n))$, for any string v of length n. However, nothing is known about the tightness of this upper bound. We present infinite families of binary strings for which $\rho(v) = \Theta(\log n)$ holds, thus giving the first non-trivial lower bound on $\rho(n)$, the maximum over all strings of length n.

Our results suggest that r is not an ideal measure of the repetitiveness of the string, since the number of repeated factors is invariant between the string and its reverse. We believe that there is a more intricate relationship between the number of runs of the BWT and the string's combinatorial properties.

The paper is available at: CoRR abs/2008.08506