An r-index for Lower Repetitiveness Scenarios

Dustin Cobas[a,b], Travis Gagie[c], Gonzalo Navarro[a,b]

[a] *CeBiB - Center of Biotechnology and Bioengineering*
[b] *Department of Computer Science, University of Chile, Chile*
[c] *Faculty of Computer Science, Dalhousie University, Canada*

Abstract

The *Run-Length FM-index* or `RLFM-index` is a variant of the `FM-index` that takes advantage of the compressibility of $BWT[1..n]$, which is formed by r runs of equal symbols. `RLFM-index` supports the count operation in $\mathcal{O}(r)$ space. However, due to the required sampling, it needs a much larger $\mathcal{O}(n/s)$ space to support locating in time proportional to s.

r`-index` is an evolution of `RLFM-index` capable to efficiently locate the occurrences of a pattern using only $\mathcal{O}(r)$ space. The experiments show that the r`-index` outperforms all the other implemented indexes by orders of magnitude in space or in time to locate pattern occurrences on highly repetitive datasets.

r is usually a relatively small number in repetitive collections ($r \ll n$). Unfortunately, the space required by r`-index` can degrade quickly when repetitiveness decreases.

In this work, we propose a new sampling mechanism for r`-index` to overcome this problem. Our approach creates a kind of hybrid sampling scheme that uses the r`-index`'s sampling over large BWT runs but, in oversampled areas of the text, it samples at regularly spaced text positions (like classic `FM-index`). Thus, we can handle areas with higher and lower repetitiveness in different ways. Our preliminar results are very promising, achieving a good space-time tradeoff.