zbMATH Open
THE FIRST RESOURCE FOR MATHEMATICS

Hooshmand, Sahar; **Tavakoli, Neda**; **Abedin, Paniz**; **Thankachan, Sharma V.**
On computing average common substring over run length encoded sequences. (English)
Zbl 1403.68373
Fundam. Inform. 163, No. 3, 267-273 (2018).

Summary: The Average Common Substring (ACS) is a popular alignment-free distance measure for phylogeny reconstruction. The ACS of a sequence $\mathrm{X}[1,x]$ w.r.t. another sequence $\mathrm{Y}[1,y]$ is

$$\mathrm{ACS}(\mathrm{X},\mathrm{Y}) = \frac{1}{x}\sum_{i=1}^{x} \max_{j} \ \mathrm{lcp}(\mathrm{X}[i,x],\mathrm{Y}[j,y])$$

The $\mathrm{lcp}(\cdot,\cdot)$ of two input sequences is the length of their longest common prefix. The ACS can be computed in $O(n)$ space and time, where $n = x + y$ is the input size. The compressed string matching is the study of string matching problems with the following twist: the input data is in a compressed format and the underling task must be performed with little or no decompression. In this paper, we revisit the ACS problem under this paradigm where the input sequences are given in their run-length encoded format. We present an algorithm to compute $\mathrm{ACS}(\mathrm{X},\mathrm{Y})$ in $O(N \log N)$ time using $O(N)$ space, where N is the total length of sequences after run-length encoding.

MSC:

68W32 Algorithms on strings
68P30 Coding and information theory (compaction, compression, models of communication, encoding schemes, etc.) (aspects in computer science)
68W40 Analysis of algorithms

Cited in **1** Document

Keywords:

string algorithms; suffix trees; RL-encoding; compression

Full Text: DOI

Edited by FIZ Karlsruhe, the European Mathematical Society and the Heidelberg Academy of Sciences and Humanities
© 2021 FIZ Karlsruhe GmbH