Journal of Automata, Languages and Combinatorics 8 (2003) 4, 647–657

AN IMPROVED ALGORITHM FOR FINDING LONGEST REPEATS WITH A MODIFIED FACTOR ORACLE [1, 2]

ARNAUD LEFEBVRE
UMR 6037 – ABISS, Université de Rouen
76821 Mont-Saint-Aignan Cedex, France
e-mail: arnaud.lefebvre@univ-rouen.fr

THIERRY LECROQ
LIFAR – ABISS, Université de Rouen
76821 Mont-Saint-Aignan Cedex, France
e-mail: thierry.lecroq@univ-rouen.fr

and

JOËL ALEXANDRE
UMR 6037 – ABISS, Université de Rouen
76821 Mont-Saint-Aignan Cedex, France
e-mail: joel.alexandre@univ-rouen.fr

ABSTRACT

We first give some experimental evidences of the difference rate on the length of the repeats of a string p found using the factor oracle of p. We show then how to improve the length of the repeats. Examples of improvements are given for finding repeats in genomic sequences and using repeats for data compression.

Keywords: String algorithms, computational biology, factor oracle, repetitions, data compression

1. Introduction

Finding repeats in strings is of great interest in areas such as bioinformatics and data compression. There exist exhaustive methods to find all the repeats in a string (see [2, 4] and [9]). The new challenge consists in dealing with huge strings such as those generated in computational biology. In [6] we introduced an on-line linear heuristic method to compute repeats in a string p using the factor oracle of p. We also showed that this method is very useful when applied on genomic sequences. However this

[1] Full version of a lecture presented at the *Thirteenth Australasian Workshop on Combinatorial Algorithms* (Kingfisher Bay Resort, Fraser Island, Queensland, Australia, July 7–10, 2002).

[2] This work was partially supported by a NATO grant PST.CLG.977017.