UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

PROJECT

Fast string alignment calculation

Kristijan Biščanić
Luka Hrabar
Ela Marušić

Zagreb, January 2016.

CONTENTS

1. Introduction

This project will implement, test and analyse a faster algorithm for computing string edit distances and sequence alignment. This algorithm was published by Masek and Paterson [3] and is inspired by the Four Russians Algorithm. Implemented algorithm will be compared to and tested against Needleman-Wunsch algorithm [4], which is based on dynamic programming.

The *string edit distance* is defined as the minimal cost of transforming one character string into the other. Operations allowed in those transformations are only insertion, deletion and replacing of one character, each of these having defined some cost. *Edit Script* is defined as the actual sequence of operations used to transform one string into the other. There are many algorithms that are using string edit distances and edit scripts for further calculations, and they are used extensively in bioinformatics for sequence alignment.

Sequence alignment is a process of arranging the symbolic representations of DNA, RNA or protein sequences so that their most similar elements are juxtaposed. Such alignment is useful to identify regions of similarity and many bioinformatics tasks depend upon successful alignments.

2. Algorithm

3. Test results

4. Conclusion

Zaključak.

5. Bibliography

[1] Dan Gusfield. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge university press, 1997.

[2] Vamsi Kundeti i Sanguthevar Rajasekaran. Extending the four russian algorithm to compute the edit script in linear space. U *Computational Science–ICCS 2008*, stranice 893–902. Springer, 2008.

[3] William J Masek i Michael S Paterson. A faster algorithm computing string edit distances. *Journal of Computer and System sciences*, 20(1):18–31, 1980.

[4] Saul B Needleman i Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

6. Abstract

Sažetak.