UNIVERSITÀ di VERONA

Scuola di Dottorato
di **SCIENZE NATURALI ED INGEGNERISTICHE**

Corso di Dottorato in Informatica

PhD Course
"Advanced Data Structures for Textual Data"

Prof. Zsuzsanna Lipták

Genome-scale textual data, i.e. strings of many giga- or even terabytes, are everywhere in today's world. This includes biological sequences (genomic data, protein sequences), digital books, web crawl data, emails, musical data, and many others. The main challenge nowadays is not so much how to store this data, but how to store it in such a way that it can be processed and queried efficiently. Text indexes are dedicated data structures for handling very large amounts of textual data. Propelled forward by the need arising from computational biology on the one hand, and from web search on the other, enormous progress has been made in this area in recent decades.

In this course, we will study some of these text indexes. We will start with a brief introduction to the suffix array, a classic data structure for strings, study its properties, some of its uses in string processing, and its efficient construction. We then introduce two supporting data structures, the LCP-array and the Burrows-Wheeler-Transform (BWT). The so-called 'clustering property' of the BWT allows compressed indexes to be built on it. We will close with several BWT-based text indexes: the FM-index, the RLFM-index, and the r-index.

Depending on the background of the students, some of the above topics may be replaced by others, such as a more thorough introduction to wavelet trees (a versatile data structure for efficient rank/select queries), or the extended Burrows-Wheeler-Transform (eBWT, a generalization of the BWT to string collections). String collections are of fundamental interest in many of the most common applications today, such as pangenomes, version control data, or web crawl data, where many different copies of highly similar strings are given in input.

Prerequisites: algorithms and data structures. The course is primarily designed for students who have no background in text indexes, but is also of interest for those who followed the masters level course "Computational Analysis of Genome-Scale Sequences", since it contains plenty of additional material.

Day 1: Introduction to problems on textual data, pattern matching, suffix arrays (SA)
Day 2: efficient SA construction, LCP-array
Day 3: BWT, backward search, wavelet trees
Day 4: FM-index, RLFM-index, r-index

total duration: 12 h (4x3 hours)

course days and times: Sept 16-19, 2024, Mo-Thu, 9:30-12:30 (lecture room to be decided)

Please sign up on the google form contained in the accompanying email by 31 August 2024.

In case of questions, please contact Prof. Lipták (email: zsuzsanna.liptak AT univr.it).