# *PhD Course "Advanced Data Structures for Textual Data"*

## *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 how to store this data, but how to store it in such a way that it can be processed and queried efficiently. Thus, dedicated data structures (referred to as *text indexes*) are necessary to handle this deluge 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 start with a brief introduction to the suffix array, a classic data structure for strings, studying 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, and present some applications in string processing. This includes a brief introduction to wavelet trees, a versatile data structure for efficient rank/select queries. Next, we will have a look at the FM-index, and, if time allows, at the r-index. Finally, we will see strategies how to handle string collections (as opposed to individual strings), in particular using the extended Burrows-Wheeler-Transform (eBWT). String collections are of fundamental interest in many of today's most common applications, 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 interesting 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 suffix arrays (SA), pattern matching on SA, efficient SA construction.
Day 2: LCP-array, Wavelet trees.
Day 3: BWT, backward search, FM-index.
Day 4: Strategies for handling string collections, eBWT, efficient eBWT construction.

**total duration**: 12 h (4x3 hours)

**course days and times**: July 11-14, 2022, 9:30-12:30 (lecture room to be decided)

In case of questions, please contact Prof. Lipták at the email address zsuzsanna.liptak@univr.it