UNIVERSITÀ di VERONA

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

Corso di Dottorato in Informatica

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 other types. The main challenge is not how to store this data, but how to store it in such a way that it can be processed and queried efficiently. Due to its sheer size, dedicated data structures (called 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 the more advanced text indices. We start with a brief introduction to the suffix array, a classic data structure for strings. We then study compressed suffix arrays, the Burrows-Wheeler-Transform, and the FM-index. We will spend some time on the combinatorics underlying the Burrows-Wheeler-Transform and suffix arrays. In the final part, we will look at data structures for sets of strings, which are especially relevant for computational biology.

Class Meetings (updated)

total duration: 20 h

course days: 21-25 September 2020

course times: Monday to Friday, 2 hours in the morning and 2 hours in the afternoon
(exact times to be agreed with the participants)

Please sign up by writing to Laura Marcazzan at: laura.marcazzan@univr.it (if not already done).

Scuola di Dottorato di Scienze Naturali ed Ingegneristiche
Strada Le Grazie, 15 - 37129 Verona | T + 045 802 7026
laura.marcazzan@univr.it
P. IVA 01541040232 | C.F. 93009870234