Università
Ca'Foscari
Venezia

PROJECT ACRONYM AND TITLE: **REGINDEX - Compressed Indexes for Regular Languages with Applications to Computational Pangenomics**

FUNDING PROGRAMME: **Horizon Europe**

CALL: **H2020 ERC - Starting Grant**

HOST DEPARTMENT: **Department of Environmental Sciences, Informatics and Statistics**

SCIENTIFIC RESPONSIBLE: **Nicola Prezza**

FINANCIAL DATA:

Project total costs	Overall funding assigned to UNIVE
€1.385.743,00	€1.385.743,00

ABSTRACT:
Sorting is, arguably, the most powerful algorithmic primitive when it comes to indexing data. At the same time, the regularities exposed by sorting are precisely those enabling data compression. In the last two decades, this fascinating duality has led researchers to the design of compressed full-text indexes: data structures supporting fast pattern matching queries over compressed text. In this project, we revisit the natural generalization of the problem to labeled graphs from a new perspective: we interpret graphs as finitestate automata and investigate the connections existing between their propensity to be sorted and the languages they recognize. Our novel language-theoretic approach makes it possible to transfer fundamental results between the mature fields of regular language theory and compressed text indexing. We aim at building this bridge by developing a new theory of compressed regular language indexing. This project finds fundamental applications to the rapidly-expanding field of computational pan-genomics, where the goal is to study the variations contained in the genomes of an entire population. Recent research has shown that representing pan-genomes as labeled graphs is an important step to reduce reference allele bias. Existing approaches, however, can index only restricted classes of graphs, thereby limiting the practical applicability of such powerful pan-genome representations. Our innovative approach, based on sorting regular languages by partial co-lexicographic orders, changes the perspective from which the compressed indexing problem has been tackled in the literature. This project aims at developing a theory of graph indexing and compression based on the natural interplay between sorting and regular language theory. We will apply these findings inside practical tools for aligning arbitrarily-long DNA fragments against compressed pan-genome graphs.

Planned Start date	Planned End date
1st September 2022	31st August 2027

PARTNERSHIP:

1	Università Ca' Foscari	Venice (IT)	Coordinator