UNIVERSITÀ DI TRENTO
Dipartimento di Matematica

DOTTORATO

CYCLE 33th
ORAL DEFENCE OF THE PHD THESIS

Friday 11 February 2022 – at 9.50 am

The event will take place online through the ZOOM platform.
To get the access codes please contact the secretary office

Elia Carlo Zirondelli

PhD Student in Mathematics

Omnitig Listing and Contig Assembly for Genomic De Bruijn Graphs

Abstract:
Genome assembly asks to reconstruct an unknown string from many shorter substrings of it. Its hardness stems both from practical issues (size and errors of real data), and from the fact that problem formulations inherently admit multiple solutions.
Given these, at their core, most state-of-the-art assemblers are based on finding non-branching paths (unitigs) in an assembly graph. If one defines a genome assembly solution as a closed arc-covering walk of the graph, then unitigs appear in all solutions, being thus safe partial solutions. All such safe walks were recently characterized as omnitigs, leading to the first safe and complete genome assembly algorithm. Even if omnitig finding was improved to quadratic time, it remained open whether the crucial linear-time feature of finding unitigs can be attained with omnitigs.
We describe an O(m)-time algorithm to identify all maximal omnitigs of a graph with n nodes and m arcs, notwithstanding the existence of families of graphs with Θ(mn) total maximal omnitig size. This is based on the discovery of a family of walks (macrotigs) with the property that all the non-trivial omnitigs are univocal extensions of subwalks of a macrotig, with two consequences: a linear-time output-sensitive algorithm enumerating all maximal omnitigs and a compact O(m) representation of all maximal omnitigs.
This safe and complete genome assembly algorithm was followed by other works improving the time bounds, as well as extending the results for different notions of assembly solution. But it remained open whether one can be complete also for models of genome assembly of practical applicability.
In this dissertation, we also present a universal framework for obtaining safe and complete algorithms which unify the previous results, while also allowing to characterize different assembly problems. This is based on a novel graph structure, called the hydrostructure of a walk, which highlights the reachability properties of the graph from the perspective of the walk. Almost all of our characterizations are directly adaptable to optimal verification algorithms, and simple enumeration algorithms. Most of these algorithms are also improved to optimality using an incremental computation procedure and a previous optimal algorithm of a specific model.

Supervisor: Romeo Rizzi

CONTATTI
Staff di Dipartimento - Matematica
tel. 0461 281508-1625-1701-3786
phd.maths@unitn.it
www.maths.unitn.it