Abstract

This dissertation thesis focuses on bioinformatic problems at multiple levels, from read mapping through tandem repeats genotyping to analysing data on the population level.

In case of read mapping, we devoted to short reads and mapping based on indexing data structures. Two new variants of the FM-index have been proposed, aiming to optimize the use of cache memory when querying the index. We managed to achieve this by reorganizing data structures of the index into aggregated independent blocks, where memory efficient storage of these structures was achieved due to the focus on the DNA and RNA alphabet.

When addressing the problem of genotyping tandem repeats, we proposed a novel tool - WarpSTR - characterizing tandem repeats directly from raw signals acquired by nanopore sequencing. WarpSTR addresses the challenges of nanopore sequencing data - high noise rate and distortion in the time domain - by representing the structure of tandem repeats in the finite-state automaton and its subsequent alignment with raw signals. To obtain the final genotype from individual estimations, we proposed a heuristic based on Gaussian mixed models. The contribution of WarpSTR was twofold. First, we significantly increased the accuracy of the characterization of simple repeats. Second, we allowed characterizing even complex tandem repeats.

In population-level analysis, we focused on epistasis detection, where epistases are combinations of variants whose interaction influences the phenotype. Nature-inspired algorithms are commonly employed to search the enormous space of all possible variant combinations. In this work, we experimented with the bat algorithm and the flower pollination algorithm, using multi-objective optimization.

Keywords: read mapping, alignment, epistases, tandem repeats, nanopore sequencing