Abstract

In this thesis, we deal with the issues of handling massive datasets in the context of computational pangenomics. The goal of pangenomics is to replace the traditional reference genome in DNA analyses with a collection of genomes from the same species, or from a set of related species. However, the analyses for large collections of genomes are computationally challenging due to the amount of data being processed. In this work, we introduce novel algorithms and data structures for handling pangenomic datasets. First, we show several results related to converting from one compression format (a straight-line program) to another (LZ77) without the need of decompressing, and in some cases in time proportional to the size of the compressed input and output. The second contribution is a novel index data structure for large pangenome graphs enabling the exact matching of whole patterns or their parts in compressed space without inducing spurious matches. The final contribution improves the time complexity of finding maximal-exact matches between a pattern and the indexed text. Inspired by this finding, we also present a heuristic that speeds up finding long matches in practice.

Keywords: compression, algorithms, pangenomics, text indexing, repetitive data