Abstract

The increasing availability of related genomic data necessitates a paradigm shift from bioinformatics analyses based on single linear reference genomes to analyses utilizing comprehensive pangenomic references. The pangenomic datasets can be characterized by two primary attributes: their vast sizes and high repetitiveness. The repetitiveness, in particular, aids in constructing pangenomic references by managing the large data sizes through two emerging approaches: pangenomic graphs and compressed stringology data structures. While pangenomic graphs combine related genomic regions into single nodes, stringology data structures utilize compression techniques directly to the textual representation of pangenome. Each of these approaches offers distinct computational, visualization, and interpretation benefits. This thesis proposes data structures that bridge the two approaches to enable transitions between graph and stringology representations.

The central data structure, introduced in Chapter 3, is named prefix-free graph. These graphs offer a scalable construction algorithm that avoids computationally expensive steps, such as multiple sequence alignment, while being flexible enough to construct graphs close to other arbitrary pangenomic graphs. Chapter 4 introduces PFG positions, enabling iteration through the suffixes of a pangenome, connecting prefix-free graphs to stringology data structures. Chapter 5 demonstrates the practical application of this connection by increasing the scalability of Wheeler graphs, a recent generalization of the Burrows-Wheeler transform, potentially unlocking the possibility of its use as a pangenomic index for datasets as large as the 1000 Genomes project. Lastly, Chapter 6 introduces the tag array, a novel data structure that relays the results of stringology methods from compressed space onto arbitrary graphs while avoiding the potentially large explosion in the number of results during decompression.

These novel data structures aim to simplify the integration of pangenomic representations, encourage the development of efficient tools for pangenomic analysis and interpretation of results, and ultimately promote broader adoption of pangenomics in biological research.

Keywords: pangenomics, pangenomic graphs, stringology, data structures, bioinformatics