Towards an edit distance between pangenome graphs

Abstract

A pangenome graph is a sequence graph that aims to represent variations among a collection of genomes in a single data structure. Each genome is segmented and embedded as a path in the graph with its successive nodes corresponding to contiguous segments on the associated genome. Building such graphs relies on alignment heuristics, and thus gives different graphs from the same input data depending on the chosen method, or the set of parameters. In this work, we would like to question to what extent the construction method influences the resulting graph and therefore to what extent the resulting graph reflects genuine genomic variations.
We present here an algorithm that analyzes the differences in segmentation across two pangenome graphs, the segmentation being the way the genomes are split into nodes inside the graph structure. We define elementary operations, fusion and fission, that enables to transform one graph into another. Our algorithm provides a dissimilarity measure between each pair of variation graphs: the minimal number of elementary operations. It enables both to quantify the impact of the graph construction method and its parameters and to pinpoint specific areas of the graph and genomes that are impacted by the changes in segmentation.
We applied our method on graphs from 21 yeast telomere-to-telomere phased genomes assemblies with the two current state-of-the-art pangenome graph builders, minigraph-cactus and pggb. We show that, with a fixed set of genomes, changing the reference in minigraph-cactus mattered much more than shuffling the order of insertion of the other genomes, and that comparing two minigraph-cactus graphs with different references can result in a higher dissimilarity than comparing a minigraph-cactus graph and the pggb graph.

Keywords: variation graph, pangenome, edit distance