Towards an edit distance between pangenome graphs

Siegfried Dubois, Benjamin Linard, Matthias Zytnicki, Claire Lemaitre, Thomas Faraut

Abstract

SeqBIM

Towards an edit distance between pangenome graphs

Siegfried Dubois[1]*, Benjamin Linard[2], Matthias Zytnicki[2], Claire Lemaitre[1] and Thomas Faraut[3]

[1]*Univ Rennes, Inria, CNRS, IRISA, Rennes, F-35000, France*
[2]*MIAT, Université de Toulouse, INRAE, 31320 Castanet-Tolosan, France*
[3]*GenPhySE, Université de Toulouse, INRAE, ENVT, 31320 Castanet-Tolosan, France*
***Corresponding author**: siegfried.dubois@inria.fr

Abstract

A variation graph is a data structure that aims to represent variations among a collection of genomes. It is a sequence graph where each genome is embedded as a path in the graph with the successive nodes, along the path, corresponding to successive segments on the associated genome sequence. Shared subpaths correspond to shared genomic regions between the genomes and divergent path to variations: this structure features inversions, insertions, deletions and substitutions. The construction of a variation graph from a collection of chromosome-size genome sequences is a difficult task that is generally adressed using a number of heuristics such as those implemented in the *state-of-the-art* pangenome graph builders *minigraph-cactus* [1] and *pggb* [2]. The question that arises is to what extent the construction method influences the resulting graph and therefore to what extent the resulting graph reflects genuine genomic variations.

We propose to address this question by constructing an edition script between two variation graphs built from the same set of genomes which provides a measure of similarity, and more importantly that enables to identify discordant regions between the two graphs. We proceed by comparing, for each genome, the two corresponding paths in the two graphs which correspond to two possibly different segmentations of the same genomic sequence. As such, for each interval defined by the nodes of the path of the genome in the first graph, we define a set of relations with the nodes of the second graph, such as equalities, prefix and suffix overlaps... which allows for a calculation of how many elementary operations, such as fusions and divisions of nodes, are required to go from one graph to another.

We tested our method on variation graphs constructed using both simulated dataset as well as a real dataset made of 15 yeast telomere-to-telomere phased genome assemblies [3], with *minigraph-cactus* as the graph construction tool. This tool builds iteratively the variation graph, starting from a genome taken as a reference and incorporating each genome in the order provided by the user. In this work, we compared by pairs the graphs constructed from the same set of genomes but using different incorporation orders. After the application of our algorithm, we get a measure of the similarity between each pair of variation graphs in the form of a

distance, that enables both to quantify the impact of the order of genomes in the graph construction, and to pinpoint the specific areas of the graph and genomes that are impacted by the changes in segmentation. Ongoing work includes being able to compare graphs issued from different variation graphs construction tools.

Availability: This algorithm is implemented as a Python tool: https://github.com/Tharos-ux/pancat

References

[1] Glenn Hickey, Jean Monlong, Jana Ebler, Adam Novak, Jordan M. Eizenga, Yan Gao, Human Pangenome Reference Consortium, Tobias Marschall, Heng Li, and Benedict Paten. Pangenome Graph Construction from Genome Alignment with Minigraph-Cactus, April 2023.

[2] Erik Garrison and Andrea Guarracino. Unbiased pangenome graphs. *Bioinformatics*, 39(1):btac743, January 2023.

[3] Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in Saccharomyces cerevisiae | Nature Genetics.