A graph-based approach to explore local structures in genome graphs aiming the identification of genetic variants

Rodrigo Theodoro Rocha[1], Georgios Joannis Pappas Junior[1]

1 DEPARTAMENTO DE BIOLOGIA CELULAR, INSTITUTO DE CIÊNCIAS BIOLÓGICAS, UNB

Abstract

Advances in DNA sequencing technologies charter the possibility to generate hundreds of genomes to capture the individual genetic variability within a population. Currently, in order to compare these, a reference genome is used as a proxy to contrast the individuals. However, a single reference falls short in representing the wealth of genetic variation. In recent years, new graph based data structures were developed to capture sequence polymorphisms along a set of genomes, collectively named genome graphs. The rationale is to reduce bias and improve biological inferences from a one-dimensional universal reference (i.e. linear sequence representation) to a multi-dimensional (i.e. genome graphs) representation of multiple genomes. This is not a trivial change of perspective and challenges common tasks in bioinformatics, including read mapping and variant detection. Borrowing concepts from graph theory, specially those concerned with connectivity of graphs, we developed a framework to identify local graph structures (motifs) that represent sequence variability sites in genome graphs. These motifs can be simple genetic variants (i.e. bubbles) or more complex structures (i.e. super-bubbles). The strategy is based on the genome graph decomposition into biconnected components, and those into triconnected components that have specific characteristics. A graph is said to be biconnected (triconnected) if it has no set of 1-vertices (2-vertices) whose removal increases the number of connected components (i.e. splits the graph). We aimed at decomposing a genome graph with respect to its triconnected components with the aid of a data structure named SPQR-tree. This can be done in linear time and we show that some graph motifs are embedded in a subset of nodes of the SPQR-tree permitting the identification of genetic variants, whereas the paths connecting the nodes therein entail the allelic variants. Moreover, given that we can sort the identified graph motifs by complexity, it is possible to correlate these to hotspots of genetic variation in the genome graph. We expect that this approach will help to interrogate pangenomes to collectively identify hyper variable sites or genomic regions under selective pressure.

Funding: Programa de Pós-graduação em Biologia Molecular, UnB e CAPES