Connection between Max-Flow and Linearization of Genome Sequence Graphs

Marija Mihova, Ilinka Ivanoska, Kire Trivodaliev

Ss. Cyril an Methodius University, Faculty of Natural Sciencies and Mathematics,
1000 Skopje, Macedonia

One of the ways to represent human genetic variation into the reference human genome is using a Genome Sequence Graph (GSG). Each node of a GSG represents a single DNA base that occurs at an orthologous locus in one or more of the haploid genomes represented. Each arc corresponds to an adjacency that occurs between consecutive instances of bases in the represented genomes, and it is directed according to the default strand direction of the DNA sequence. We can add weight to each arc in order to emphasize the importance of an arc in typical applications running on the graph, for example number of times that the arc is traversed in the reference genomes used to build the graph. The representation of the GSG itself is very important since the efficiency and the effectiveness of the operations such as access, traversal and visualization depend on that. The best representation is ordering the nodes in a straight line, which is known as linearization of the graph. Usually the linearization of a sequence graph aims to lessen the total weight of all feedback arcs, i.e. the weighted feedback, as much as possible. We can modify the sequence graph to obtain a flow network in the following manner. We will add special source and a special sink to the set of nodes. For each node that represents a starting DNA base, we add an arc from the source to it, with a weight equal to the number of genomes that start with it. Similarly, for each node that represents an ending DNA base, we add a weighted arc from that node to the sink. The weight of that arc is equal to the number of genomes that end with that node. With this modification of the graph the obtained weight function is a flow function as well, because the total flow out of the source is equal to the total flow entering the sink, and for all other nodes, total flow leaving a node is equal to the total flow entering a node [2, 3]. The max-flow, M, of this network is equal to the total flow out of the source, so any minimal path vector for level M, i.e. minimal flow function for flow M, is an undirected graph. The features of these minimal paths can help in minimizing the weighted feedback of the graph. In this paper we give some theoretical results that lead to design of an algorithm for reducing feedback edges in a linearized graph. This way of representation of a directed weighted graph is important for representation of Genome Sequence Graphs, since it improves the efficiency and the effectiveness of the operations such as access, traversal and visualization of the sequences. We will continue our work in this direction, and our goal is to design an algorithm that produces the best linearization in this sense.