Application of Computational Geometry Algorithms in Identification of Structural Variations in a Genome Sequence

Identifying structural variations is a problem in bioinformatics to measure the differences between the DNA sequences of two organisms (Feuk et al. 2006). A structural variation can be insertion, deletion, inversion and duplication of a sequence. Sindi et al. (2009) describes a methodology to identify the structural variants as a computational geometry problem. For a given alignment of a segment of DNA from the test genome to the reference genome, they derive a polygonal area in which a break point of a structural variation is likely to begin to occur. The shape of these break point regions can be either trapezoids or rectangles. Then by finding the intersections of these polygons, and comparing different DNA fragments belonging to the same area, the region corresponding the break point can be refined. To solve this problem they define an efficient plane sweep algorithm to find the intersections of these polygons.

The aim of this project is to implement a polygon intersection algorithm that is suitable for finding such intersections. The algorithm should be able find all non-empty intersection regions resulting from all possible subsets of the polygonal break point regions. It should be also capable of reporting the region with the maximal number of polygon intersections if more than two polygons are intersecting at once. Thus, clusters of polygons with similar properties can be also identified.

References
