OPTIMAL PERIODIC CLUSTERING

Tathagata Debnath[1], and Mingzhou Song[1,2]

[1]Department of Computer Science, [2]Molecular Biology and Interdisciplinary Life Sciences Graduate Program, New Mexico State University

Contact email: tathadbn@nmsu.edu, joemsong@nmsu.edu

Periodic data are abundantly present in real life such as circadian rhythm in animals and plants, year around wind direction data and ocean current data, solar spot activity data etc. In genetics, we have circular data such as the bacterial genome and plant chloroplast. Clustering gene locations in those genomes present a particular challenge compounded by the lack of proper starting position. This results in a runtime of $O(N^2t)$ for the iterative *K*-means algorithm, where N is the number of points in the data and t is the number of iterations taken by the *K*-means algorithm. Moreover, the clustering outcome generated by the *K*-means algorithm is not guaranteed to be optimal. To address all these issues, we have developed the Fast Optimal Circular Clustering Algorithm (FOCC). The results of FOCC algorithm are guaranteed to be optimal. Moreover, the runtime of the FOCC algorithm is poly-logarithmic, resulting in faster execution time for large datasets. FOCC exabits superior performance than the heuristic *K*-means algorithm for simulated as well as real circular data, thus filling the void in literature. We hope the FOCC algorithm will help the research community unveil many interesting scientific phenomena in the periodic data.