KMD Clustering: Robust Generic Clustering of Biological Data

Aviv Zelig[1,2] and Noam Kaplan[2]

[1] - Data Science & Engineering Program, Faculty of Industrial Engineering & Management, Technion - Israel Institute of Technology

[2] - Department of Physiology, Biophysics & Systems Biology, Rappaport Faculty of Medicine, Technion – Israel Institute of Technology

The challenges of clustering noisy high-dimensional biological data have spawned clustering algorithms that are tailored for specific subtypes of biological datatypes. However, the performance of such methods varies greatly between datasets, they require post hoc tuning of cryptic hyperparameters, and they are often not transferable to other types of data. Here we present a novel generic clustering approach called KMD clustering, based on a simple generalization of single and average linkage hierarchical clustering. We show how a generalized silhouette-like function is predictive of clustering accuracy and exploit this to eliminate the main hyperparameter. We evaluated KMD clustering on high-noise simulated datasets, mass cytometry datasets and scRNA-seq datasets. We also implemented a sampling-based approach to extend our method to large datasets of one million cells. Compared to generic and state-of-the-art specialized algorithms, KMD clustering's performance was better or comparable to that of the best algorithm on each of the tested datasets.