Clustering Methods: from k-means to Gaussian mixture model and Louvain algorithm

In this seminar series, I will review some basic concepts of clustering analyses that are commonly used in cancer research. I will talk about three types of clustering analysis methods: 1) traditional methods of hierarchical clustering and k-means; 2) Gaussian mixture model (GMM) and latent Dirichlet allocation (LDA); 3) graph-based approaches, which are the state-of-the-art technologies used in the single cell data analysis. Hierarchical clustering builds a dendrogram by sequentially merging samples that have the shortest distance. When two samples are joined together, they form a new internal node. The process continues until all samples are merged and forms a dendrogram. The resulting dendrogram is a tree with the samples represented by the leaf nodes and the tree can be cut at different heights to generate clusters of samples. Hierarchical clustering is a bottom-up approach. In contrast, k-means is top-down approach. Samples are partitioned into k clusters. The objective function of k-means is the sum of squared distances between sample and its centroid in the cluster. Solution of k-means can be obtained through expectation maximization (EM) algorithm. Gaussian mixture model (GMM) can be thought of as an extension to k-means. Instead of having hard assignment, GMM generates probability distribution of a sample to each cluster as soft assignment. I will also introduce non-negative matrix factorization (NMF), which is dimension reduction method and explain the connection between the clustering analysis and dimension reduction methods. One can reformulate k-means method with matrix transformation and it can be shown that k-means clustering is equivalent to sparse NMF. I will also explain the connection between latent Dirichlet allocation (LDA) and NMF as well as GMM. The k-means clustering is effective if clusters are linearly separable. However, if samples are located in non-Euclidean space, we need non-linear methods. The two commonly used non-linear methods are graph-based approaches and kernel methods. I will talk about spectral clustering, which is a graph-based method and consists of dimension reduction with Laplacian Eigenmap and k-means clustering in the reduced dimension space. I will also talk about Louvain algorithm, which is used in Seurat package to cluster single cell RNAseq data. Louvain algorithm is a network community approach. It is very fast and has capacity to do clustering analysis for million nodes in a network. I will provide practical examples to illustrate how each method works and how to interpret the results of clustering analysis and explain the pros and cons of each method.