4. Which of the clustering algorithms discussed in the lecture so far can you apply to cluster biological sequences? Give reason for your answer!

Hierarchical clustering is widely used for detecting clusters in genomic data. It generates a set of partitions forming a cluster hierarchy. According to linkage criteria, there are three hierarchical clustering methods including single-linkage clustering (SL), complete-linkage clustering (CL) and average-linkage clustering (AL). With SL, clusters may be merged together due to single sequences being close to each other, even though many of the sequences in each cluster may be very distant to each other. CL tends to find compact clusters of approximately equal diameters. With CL, all objects in a cluster are similar to each other. AL can be seen as an intermediate between single and complete linkage clustering, resulting in more homogeneous clusters than those obtained by the single-linkage method. Hierarchical approaches may yield fairly good results, but they require the similarity of all pairs of sequences and quickly arrive at a bottleneck in terms of computational time and memory usage for large-scale data sets.

On the other side, K-means (KM) is a commonly used method of partitional clustering methods. KM has a lower order of computational complexity and demands less physical memory than the hierarchical method. It is suitable for clustering large gene data. The major drawback of KM compared to hierarchical clustering algorithms is the lack of hierarchical relationships in its results.