Assignment Report for k-Means Clustering

This report presents my findings from k-means clustering analysis on dataset that consists amino acid contents of proteomes of prokaryotic genomes. C++ implementation of k-means clustering was made for this analysis. Clustering program was run with different k's from 1 to 10. For each k within cluster sum of squares (WCSS), the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) results were calculated. Figure 1 shows those results for each k. These results showed that while WCSS decreases while k increases. For BIC and AIC, the situation is different. BIC results showed more like an elbow shaped trend. It was expected for BIC and WCSS but AIC results only started to increase again after k=6. It has been also observed that even though overall trend was as expected, there were some unexpected peaks as you can see in the Figure 1. It's my assumption that it was because my program was not be able to choose the best centroids when it's iterating or it's not getting the minimum WCSS when it is iterating through different k's.

Second, it has also been observed that while k is increasing species tends to cluster together. I wasn't be able to observe that GC content is affecting the clustering results. It has been observed that genomes from same families were also tend to cluster together, for example Shewanella and Halobacteria. The results also showed that genomes with higher optimal growth temperature were clustered together.

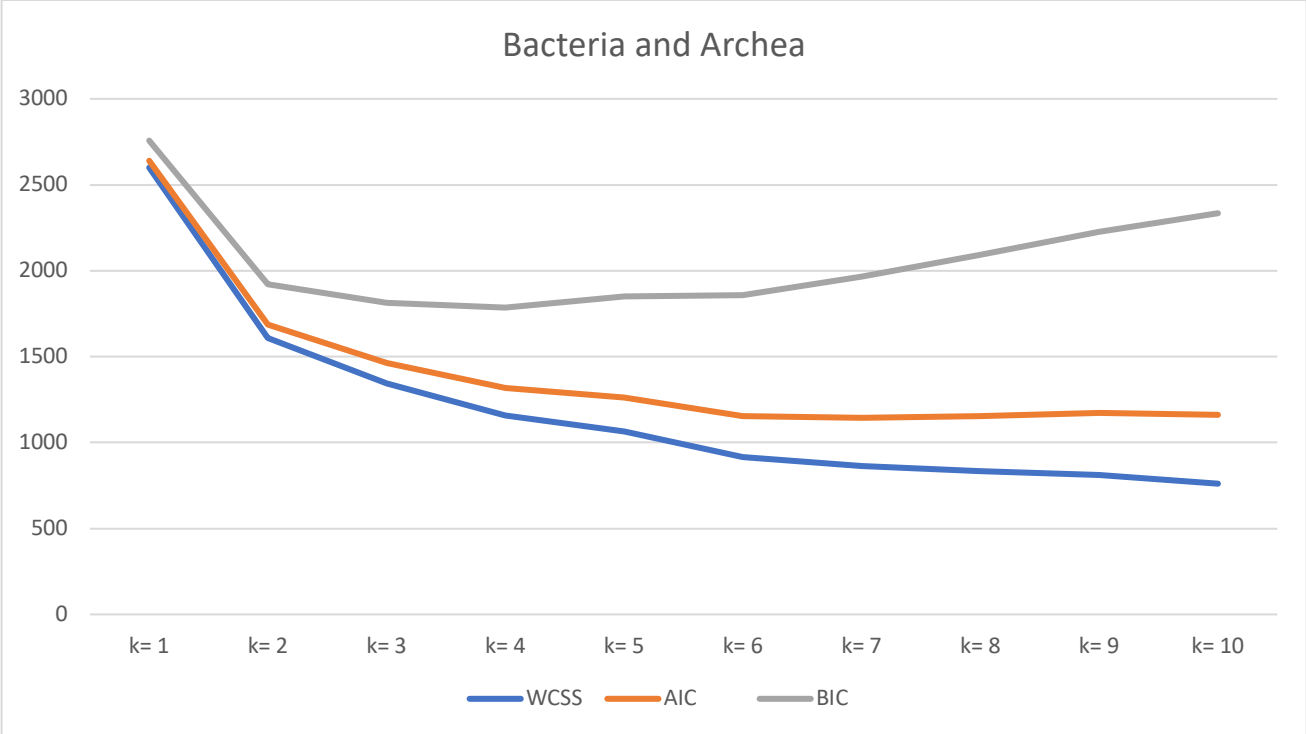

Figure1. This figure shows within cluster sum of squares (WCSS), the Akaike information criterion (AIC), the Bayesian information criterion (BIC) results for each k from k-means clustering results of Bacteria + Archea data.

k	WCSS	AIC	BIC
k= 1	2600	2640	2757.27
k= 2	1606.91	1686.91	1921.44
k= 3	1342.92	1462.92	1814.72
k= 4	1156.42	1316.42	1785.48
k= 5	1063.65	1263.65	1849.98
k= 6	914.713	1154.71	1858.3
k= 7	864.106	1144.11	1964.96
k= 8	835.487	1155.49	2093.61
k= 9	812.686	1172.69	2228.07
k= 10	760.942	1160.94	2333.6

Table1. Within cluster sum of squares (WCSS), the Akaike information criterion (AIC), the Bayesian information criterion (BIC) results for each k from k-means clustering results of Bacteria + Archea data.