Assignment 2 for BINF 8500

Summary table for Bacteria and Archaea K-means result:

Table 1. Summary for K-means results

k	MeanDisance	WCSS	AIC	BIC
1	4.4379	2580.0000	2620.0000	2677.3507
2	3.4886	1594.2814	1674.2814	1788.9828
3	3.1870	1330.5949	1450.5949	1622.6469
4	2.9403	1132.5120	1292.5120	1521.9147
5	2.7000	954.9834	1154.9834	1441.7369
6	2.5836	874.4281	1114.4281	1458.5322
7	2.4647	795.8002	1075.8002	1477.2550
8	2.3503	723.6439	1043.6439	1502.4495
9	2.2577	667.7373	1027.7373	1543.8935
10	2.1716	617.7568	1017.7568	1591.2637
11	2.0852	569.5979	1009.5979	1640.4555
12	2.0150	531.9053	1011.9053	1700.1136
13	1.9687	507.7385	1027.7385	1773.2974
14	1.9182	481.9885	1041.9885	1844.8982
15	1.8711	458.6482	1058.6482	1918.9085

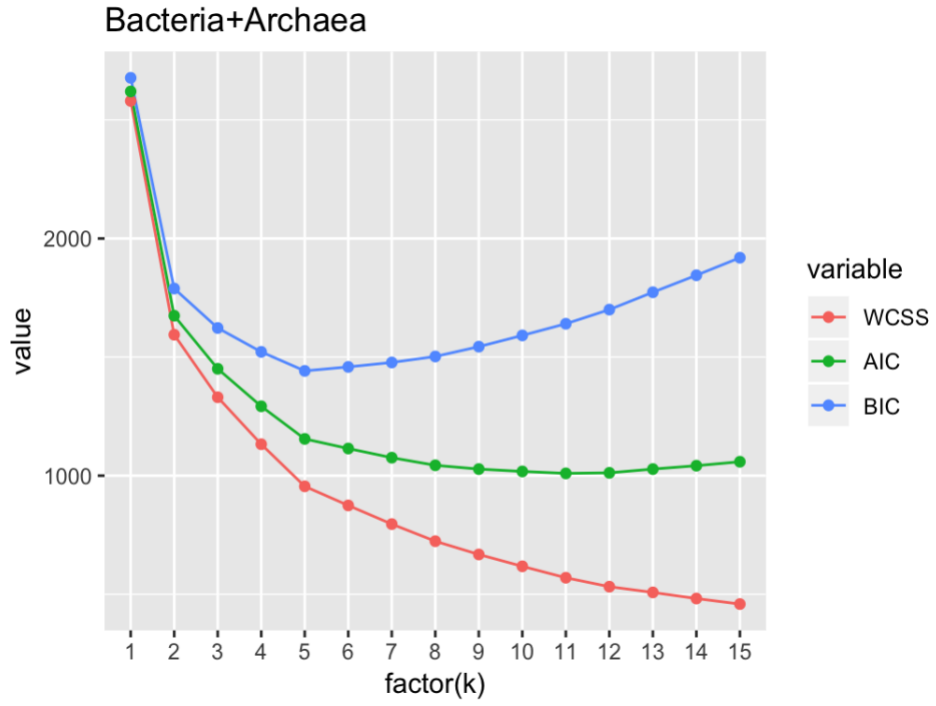

Figure 1. Visualized summary for K-means results. The lowest BIC locates at k=5.

Figure 2 shows the GC content and optimal growth temperature for each species, annotated with the cluster results from amino acid' coordinates. Species in the 1st cluster have high GC content (around 60%-70%). Their optimal growth temperatures are variant, but most of them are relatively low. Species in the 2nd cluster have low to medium optimal growth temperature, and the GC contents are not far away from 50%, all of them under 60%. For the 3rd cluster, species have low optimal growth temperature (less than 45) and high GC contents around 60%. Species in the 4th cluster have low GC contents, mostly less than 40%, with large range of optimal growth temperatures. In the 5th cluster, species are mostly heat-resistant, with relatively high optimal growth temperatures. Their GC contents are medium, around 40% to 55%.

The clustering result reveals that both GC contents and optimal growth temperature could influence the amino acid preference. In addition, when the GC content is near 50%, optimal growth temperature tends to become the dominant factor for amino acid preference (Cluster 2 and Cluster 5). Otherwise, when the GC content is relatively far away from 50%, it becomes the most significant factor for amino acid preference (Cluster 4 and Cluster 1/3).

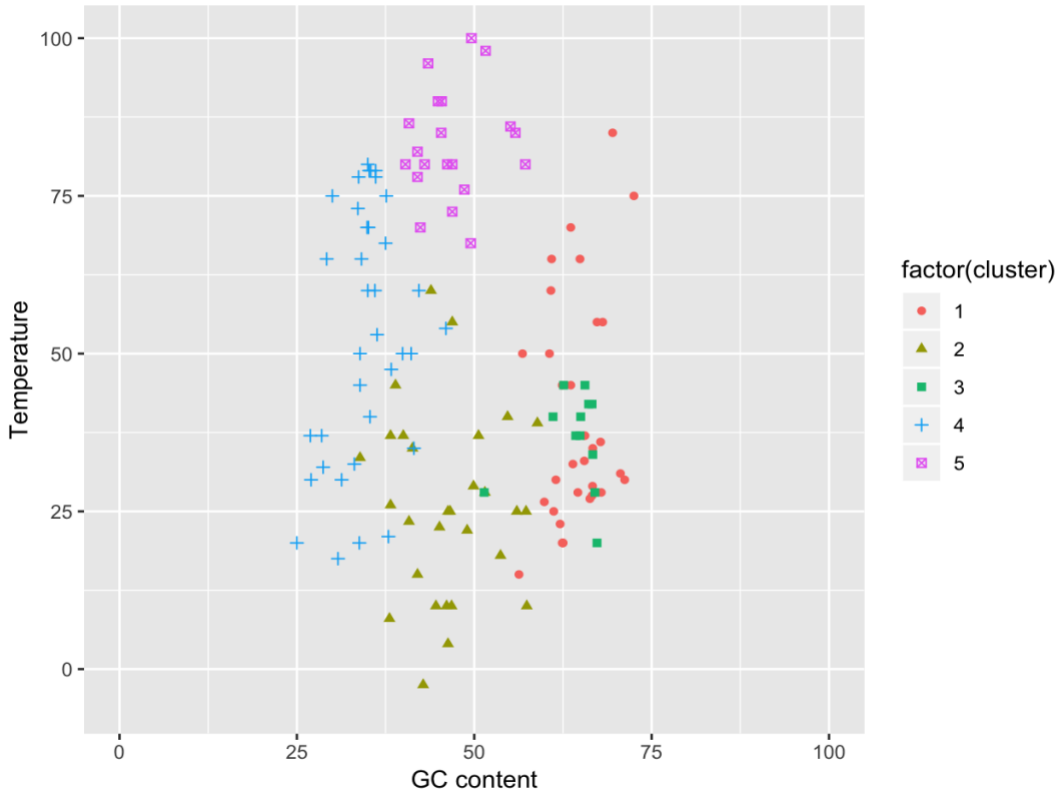

Figure 2. The dot plot showing clusters from amino acid preference are related to GC content and growth temperature