BINF 8500 Assignment2--Report for k-means cluster

Jiani Chen



|  |  |  |  |
| --- | --- | --- | --- |
| **k** | **WCSS** | **AIC** | **BIC** |
| 1 | 2600 | 2640 | 2697.35069 |
| 2 | 1606.64016 | 1686.64016 | 1801.34154 |
| 3 | 1340.90956 | 1460.90956 | 1632.96163 |
| 4 | 1141.29112 | 1301.29112 | 1530.69387 |
| 5 | 963.091967 | 1163.09197 | 1449.84541 |
| 6 | 882.517252 | 1122.51725 | 1466.62139 |
| 7 | 804.35438 | 1084.35438 | 1485.8092 |
| 8 | 730.095479 | 1050.09548 | 1508.90099 |
| 9 | 678.909326 | 1038.90933 | 1555.06553 |
| 10 | 628.545276 | 1028.54528 | 1602.05217 |
| 11 | 579.006052 | 1019.00605 | 1649.86363 |
| 12 | 540.566929 | 1020.56693 | 1708.7752 |
| 13 | 523.817564 | 1043.81756 | 1789.37652 |
| 14 | 497.209956 | 1057.20996 | 1860.1196 |

For the data “Bacteria+Archaea.txt”, the best k is 5 using Bayesian information criterion. I set configurations time as 300 in my code. Since I get a smooth line of WCSS.AIC.BIC, which suggest that 300 time is enough to find a reasonable global minimal WCSS for k=1-14.



Using k=5 that I find above, the genome was clustered as above and plot with two principle factors. Generally, the genome that are clustered by my program were clustered together based on those two factors, which suggested the amino acid components of genomes were closely related to genome GC content and optimal growth temperature.