Jack Ustanik

Chart, line chart

Description automatically generatedAssignment 3 Kmeans clustering

In different bacterial genomes percentages of amino acids have a relationship regarding optimal growth temperature and GC content. Genomes which contain similar percentages of each amino acid will have similar optimal growth temperatures and GC content. The purpose of the Kmeans algorithm is to group similar data points into a certain number of clusters. The Kmeans is used here to cluster prokaryotic genomes based on amino acid contents or their proteomes.

The data and visuals for my report are for the bacteria file. The number of optimal clusters using the Bayesian information criterion (BIC) and the elbow method is roughly 3. I chose to use this method to find the right number of clusters (K=3). The 1st and 3rd cluster had 28 values each and the second cluster had 35 values. The calculated BIC value was 1102.04, Akaike information criterion (AIC) was 951.39, and the within-cluster sum of squares (WCSS) value was 831.39 for 3 clusters. It was apparent that the optimal K value chosen by the BIC method was about 3 or 4 lower than the using the AIC method which was roughly 6 or 7.

In cluster 1 the optimal growth temperatures ranged from -2.5C to 60C with a mean temperature of 25.875. The GC content of this cluster ranged from 33.9% to 58.9% with an average GC content of 44.62%. In cluster 2 the optimal growth temperatures ranged from 15C to 85C with a mean temperature of 39.1C. The GC content of this cluster ranged from 51.4% to 72.5% with an average GC content of 60.93%. In cluster 3 the optimal growth temperatures ranged from 17.5C to 96C with a mean temperature of 60.77C. The GC content of this cluster ranged from 25% to 46.9% with an average GC content of 33.62%. Bacteria which shared the same genus all ended up being in the same cluster.

The distinct average temperatures, distinct percentage of GC, and similar geneses among clusters led me to believe that the clustering algorithm is working as expected. Other than some outliers the optimal growth temperatures and GC content of the clusters stayed close around the average values.

The Exact Clusters of various bacteria which I used are on the next page

|  |  |  |
| --- | --- | --- |
| Cluster # 1   1. Bacillus coagulans 2-6|55|46.9 2. Bartonella henselae Houston-1|37|38.2 3. Bdellovibrio bacteriovorus Tiberius|29|49.9 4. Colwellia psychrerythraea 34H|8|38.05 5. Desulfobulbus propionicus 1pr3, DSM 2032|39|58.9 6. Desulfotalea psychrophila LSv54|10|46.8 7. Dyadobacter fermentans NS114, DSM 18053|28|51.5 8. Enterococcus faecium Hucker1, NRRL B-2354|26|38.2 9. Escherichia coli O145:H28 RM13514|37|50.6 10. Haemophilus parasuis SH0165|37|40 11. Halothiobacillus neapolitanus c2, ATCC 23641|40|54.7 12. Lactococcus lactis subsp. cremoris NZ9000|40|35.3 13. Oenococcus oeni PSU-1|21|37.9 14. Parageobacillus thermoglucosidasius C56-YS93|60|43.9 15. Pelodictyon (Chlorobium) luteolum DSM 273|25|57.3 16. Photobacterium profundum SS9|15|42 17. Psychrobacter arcticus 273-4|-2.5|42.8 18. Runella slithyformis LSU4, DSM 19594|25|46.6 19. Serratia plymuthica AS9|25|56 20. Shewanella baltica OS223|4|46.3 21. Shewanella denitrificans OS217|22.5|45.1 22. Shewanella halifaxensis HAW-EB4|10|44.6 23. Shewanella loihica PV-4|18|53.7 24. Shewanella sediminis HAW-EB3|10|46.1 25. Staphylococcus lugdunensis N920143|33.5|33.9 26. Streptococcus thermophilus CNRZ1066|45|38.9 27. Tolumonas auensis TA 4, DSM 9187|22|49 28. Vibrio vulnificus M06-24/O|25|46.3 | Cluster # 2   1. Agrobacterium radiobacter K84|26.5|59.87 2. Azotobacter chroococcum X50Fred|27|66.3 3. Bordetella parapertussis Bpp5|36|67.8 4. Candidatus Desulforudis audaxviator MP104C|60|60.8 5. Corynebacterium maris DSM 45190|35|66.7 6. Corynebacterium terpenotabidum Y-11|28|67 7. Deinococcus actinosclerus BM2|31|70.6 8. Desulfuromonas sp. WTL|25|61.2 9. Dinoroseobacter shibae DFL-12, DSM 16493|33|65.51 10. Hyphomicrobium nitrativorans NL23|32.5|63.9 11. Isosphaera pallida IS1B, ATCC 43644|45|62.4 12. Kocuria rhizophila DC2201|30|71.2 13. Leisingera methylohalidivorans MB2, DSM 14336|20|62.4 14. lumatobacter coccineum YM16-304|20|67.3 15. Methylococcus capsulatus Bath|45|63.6 16. Mycobacterium tuberculosis H37Rv|37|65.6 17. Paenibacillus borealis DSM 13188|28|51.4 18. Polaromonas naphthalenivorans CJ2|20|62.5 19. Polaromonas sp. JS666|20|62.5 20. Pseudomonas aeruginosa PAO1|27.5|66.6 21. Pseudomonas putida S16|30|61.5 22. Renibacterium salmoninarum ATCC 33209|15|56.3 23. Saccharomonospora viridis P101, DSM 43017|55|67.3 24. Salinibacter ruber M31, DSM 13855|42|66.12 25. Sphaerobacter thermophilus 4ac11, DSM 20745|55|68.1 26. Sphingomonas sanxanigenens DSM 19645, NX02|29|66.67 27. Spirochaeta thermophila Z-1203, DSM 6578|65|60.9 28. Sulfobacillus acidophilus TPY|50|56.8 29. Tessaracoccus flavus RP1T|28|67.9 30. Thermaerobacter marianensis 7p75a, DSM 12885|75|72.5 31. Thermobacillus composti KWC4, DSM 18247|50|60.6 32. Thermomicrobium roseum DSM 515|70|63.6 33. Thermus scotoductus SA-01, ATCC 700910|65|64.9 34. Thermus thermophilus HB8|85|69.5 35. Xanthomonas citri citri Aw|28|64.6 | Cluster #3   1. Aquifex aeolicus VF5|96|43.5 2. Caldanaerobacter subterraneus tengcongensis MB4T|75|37.6 3. Caldicellulosiruptor hydrothermalis 108|79|36.1 4. Caldicellulosiruptor kristjanssonii 177R1B, DSM 12137|78|36.1 5. Caldicellulosiruptor kronotskyensis 2002|70|35.1 6. Caldicellulosiruptor obsidiansis OB47|79|35.2 7. Caldicellulosiruptor owensensis OL|79|35.4 8. Carboxydothermus hydrogenoformans Z-2901|78|42 9. Cellulophaga algicola IC166, DSM 14237|20|33.8 10. Clostridium saccharobutylicum DSM 13864|32|28.7 11. Clostridium stercorarium stercorarium DSM 8532|60|42.2 12. Desulfurobacterium thermolithotrophum BSA, DSM 11699|70|34.9 13. Dictyoglomus thermophilum H-6-12, ATCC 35947|78|33.7 14. Flexistipes sinusarabici MAS10, DSM 4947|47.5|38.3 15. Ignavibacterium album Mat9-16, JCM 16511|45|33.9 16. Lacinutrix sp. 5H-3-7-4|17.5|30.8 17. Marinitoga piezophila KA3|65|29.2 18. Mesoplasma florum L1|30|27 19. Mycoplasma hyopneumoniae 7422|37|28.5 20. Mycoplasma mobile 163K|20|25 21. Natranaerobius thermophilus JW/NM-WN-LF|53|36.3 22. Pseudothermotoga thermarum DSM 5069|80|40.3 23. Thermoanaerobacter kivui DSM 2030|60|35 24. Thermocrinis albus HI 11/12, DSM 14484|80|46.9 25. Thermodesulfatator indicus CIR29812, DSM 15286|70|42.4 26. Thermodesulfobium narugense Na82, DSM 14796|50|33.9 27. Thermotoga maritima MSB8|80|46.2 28. Thermotoga neapolitana DSM 4359|72.5|46.9 |