Jacob Needham

A01874339

CS 5050 Advanced Algos

Assignment 4 DNA Sequencing

Summary

Disclaimer:

An important caveat for this assignment is that my computer was struggling to handle the full files. It would create the cache but couldn’t write to file. The error code was “Process finished with exit code -1073741571 (0xC00000FD)” Anyway, the algo works on smaller data sets. So broke the first 1000 lines from the genes and ran a comparison against each of those for a quick test case. Given more time I would solve it for the full cases and split the problem into smaller problems to solve.

Discussion

Investigating the relative score various DNA sources

|  |  |  |
| --- | --- | --- |
| Data Set | Number of Nucleotides per DNA | Score |
| Homosapien v Neanderthal | 1000 | 4932 |
| Homosapien v Gorilla | 1000 | 2205 |
| Gorilla v Neanderthal | 1000 | 2191 |

Its really interesting that we (Homosapiens) share a good amount of DNA with Neanderthals but significantly less than with gorillas. Gorillas are also relatively removed from Neanderthals.

With the rise of the viruses like the Corona Virus, it would be extremely useful to have a way to know from which source the virus came. It would be really interesting to sample a bunch of DNA sources from which you thought the virus could have mutated from and see which source is closest.

I also wonder how the similarity chart was developed. It would be interesting to know how they scored a mutation or change in a DNA sequence.