ACGAAAAAAGACGGGAAAACCGTCTATCAATATGCTGAAACGCGTGAGAAACCGTGTGTCAACTGGATCACAGTTGGCGAAG
AGATTCTCAAGAGGATTGCTGAACGGTCAAGGACCAATGAAATTGGTTATGGCGTTCATAGCTTTCCTCAGATTTCTAGCCA

DLNA

Using Deep Learning to Classify DNA Sequences

Gabriel Gallardo, Liam O'Connor, Samuel Murk Caya, Jacques von Steuben

Introduction

We sought to improve upon the methodology of a paper by Gunasekaran et al., wherein the authors used convolutional neural networks and other hybrid models to classify the DNA sequences of MERS-CoV, SARS-CoV-1, SARS-CoV-2, dengue, hepatitis, and influenza. Likewise, we developed a convolutional neural network in order to classify the same six viruses, but we approached the preprocessing and architecture somewhat differently. Moreover, we also compared our results and that of the paper to baseline models that make classifications based upon the length of a DNA sequence and upon the guanine-cytosine (GC) content of a sequence.

We chose the paper because of its relevance and impact to the real world, especially with an ongoing pandemic. Rapid identification of pathogens is of vital importance to the health and safety of our communities.

Methodology

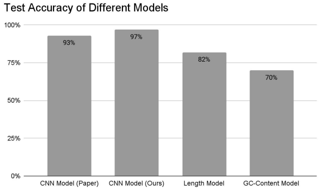
Preprocessing: We collected whole or nearly whole genomes for the six viruses from a public nucleotide database. We then randomly sampled up to 1000 DNA sequences each and then divided each sequence into subsequences of 300 nucleotides. We also converted the DNA nucleotides into one-hot encodings.

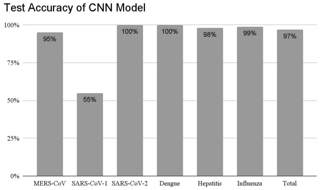
Architecture

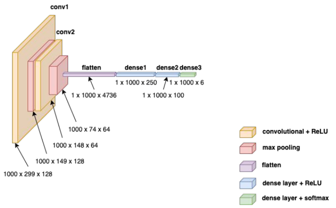
CNN Model: For our CNN model, we fed the batched subsequences through two one-dimensional convolution layers with max pooling and then three dense layers. The convolution layers and the first two dense layers use ReLU, and the third dense layer uses softmax.

Length Model: For the length baseline model, we fed the lengths of the entire DNA sequences through three dense layers, all of which do not use activation functions. Then we fed the output of the third dense layer through a softmax.

GC-Content Model: For the GC-content baseline model, we fed the GC-content ratio of the entire DNA sequences through three dense layers, all of which do not use activation functions. Then we fed the output of the third dense layer through a softmax.







Results

CNN Model: The overall accuracy of the model is 97%, but the individual accuracies for each of the six viruses are 97% for MERS-CoV, 55% for SARS-CoV-1, 100% for SARS-CoV-2, 100% for dengue, 98% for hepatitis, and 99% for influenza.

Length Model: The overall accuracy of the model is 82%, but the individual accuracies for each of the six viruses are 0% for MERS-CoV, 0% for SARS-CoV-1, 100% for SARS-CoV-2, 100% for dengue, 100% for hepatitis, and 100% for influenza.

GC-Content Model: The overall accuracy of the model is 70%, but the individual accuracies for each of the six viruses are 0% for MERS-CoV, 0% for SARS-CoV-1, 100% for SARS-CoV-2, 98.4% for dengue, 99.5% for hepatitis, and 59% for influenza.

Discussion

Lessons Learned: Sometimes simpler is better! Our much simpler models performed about as well on SARS-CoV-2, dengue, and hepatitis as our more complex CNN model.

Lingering Problems: Similar to the paper, our CNN model often misclassified SARS-CoV-1 as SARS-CoV-2. Our length model and GC content models struggled to differentiate between all three of the coronaviruses, and often classified any coronavirus sample as SARS-CoV-2, because there were more SARS-CoV-2 samples in our data.

Future Work: Had we the time, it would have been interesting to work on the interpretability of our CNN model, perhaps visualizing the filter weights of the convolution layers to see which patterns in the subsequences are important. We also could have attempted to solve the inability to different coronaviruses.

Original Paper

Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S., Venkatesan, C., & Suresh Gnana Dhas, C. (2021). Analysis of DNA Sequence Classification Using CNN and Hybrid Models. Computational and mathematical methods in medicine, 2021, 1835056. https://doi.org/10.1155/2021/1835056

TTCAAACATGCTGAGCATAATCAACAAAAGGAAAAAGACATCGCTCTGTCTCATGATGATGTTACCAGCAACACTTGCTTTC
CACTTGACTTCACGCGATGGAGAGCCGCGCATGATTGTGGGAAGAATGAAAGAGGGAAATCCTTACTTTTTAAGACAGCTT