Quantifying the Effect of Covid-19 on Gene Expression

Anna McDonald

1. Introduction

This project aims to further develop the knowledge of the human body by learning the way genes are impacted by a virus, like Covid-19. The topic of bioinformatics continues to rapidly develop new methods for the analysis of gene expression. I plan to investigate cluster analysis and classification in hopes of discovering key differences between patients who tested positive or negative for Covid-19. I hope to develop a model that could not be viewed as fitting the data just by chance.

The motivation behind this project is the answers these methods could provide to the world of health care. From things like being able to classify the presence of cancer due to gene expression to being able to predict what characteristics could lead to a type of disease, the field of health informatics aims to find the solution to things like that and more.

2. Data Mining Task

The first task is to discover how genes will cluster based off gene expression in patients who were tested for Covid-19. This will be done through multidimensional scaling analysis. I will also find what genes are differentially expressed between different treatment groups.

Next, I will create a machine learning algorithm to classify patients as either positive or negative for Covid-19 based off of the significant DE genes. I will test my algorithm with different learning rates and number of iterations. I will evaluate accuracy by taking the total number of correct identifications and dividing it by the total number of classifications made for both the training and testing data.

Questions

- Which factors play the biggest role in the grouping of gene expression?
- Can the results of a Covid-19 test be correctly classified using gene expression?

3. Methodology

All code is available here: https://github.com/annamcd511/CPTS315.git

(i) Pre-Processing

I first created a sample key for the data. Each sample had a unique combination of the following features; Individual, age, gender, infection (positive or negative), and viral load. Then the gene count matrix and the newly created sample key were loaded into R (version 4.2.1). Genes were filtered out if there were less than 0.5 counts per million in at least three individuals, resulting in 27,877 genes.

(ii) Differential Gene Expression Analysis

To visualize overall variation in gene expression, the top 10,000 variable genes were plotted on a multidimensional scaling (MDS) plot. Differential expression was quantified using the Bioconductor package edgeR (version 3.40.0), implementing a generalized linear model (GLM). To test for statistical significance, we used the GLM quasilikelihood F test (glmQLFTest in edgeR) with a Benjamini-Hochberg correction. Significance was set at a false discovery rate (FDR) <0.05. The up and down regulated genes for the comparison between the individuals that tested positive and negative for Covid-19 were found.

(iii) Binary-Classifier with Perceptron Weight Update

The 15 most differentially expressed genes (ten up regulated, five down regulated) were saved into a new csv file. This file was loaded into Python (version 3.8) and about 75% of the samples were designated to the training set whereas the remaining 100 samples were used for testing purposes. Using the pseudocode from Algorithm 1, the classifier was tested with different learning rates (n = 0.1, 1, 10) and iterations (1-20) to find the combination that will result in the highest training accuracy.

Algorithm 1 Online Binary-Classifier Learning Algorithm

Input: $\mathcal{D}$ = Training examples, T = maximum number of training iterations
Output: w, the final weight vector

Initialize the weights $w = 0$
for each training iteration $itr \in \{1, 2, \cdots, T\}$ **do**
  for each training example $(x_t, y_t) \in \mathcal{D}$ **do**
    $\hat{y}_t = sign(w \cdot x_t)$ *// predict using the current weights*
    if mistake **then**
      $w = w + \eta \cdot y_t \cdot x_t$ *// update the weights*
    end if
  end for
end for
return final weight vector w

(iv) Challenges

One challenge I came across was having to reduce the dimensionality of the data. Before finding the number of DEgenes, I expected it to be no greater than 100. However, I was left with 13,075 which was too many to use as features in my classifier. I also had a problem where my training accuracy was extremely high, but the testing accuracy was significantly lower. I found this to be due to the fact that the samples were sorted with all of the positive samples at the top and the negatives ones at the bottom. To create batches that were more representative of the overall dataset I implemented a shuffle function that I could perform before each iteration.

(v) Evaluation Methodology

Data was downloaded from Gene Expression Omnibus (GEO) repository on the National Center for Biotechnology Information website. It was originally part of a study that took place at the University of Washington [1]. The dataset consists of 484 individuals (columns) who were tested for Covid-19 and the number of reads present that correspond to each of the 35,784 genes (rows).

To ensure that I have accurate and valid results, a subset of the data was withheld from the training process for testing purposes. Additionally, the number of incorrect classifications after running the testing data through the algorithm was recorded to make sure there was not a high occurrence of false positives. I also compared my findings with the previous study to see if my results were logical.

4. Results

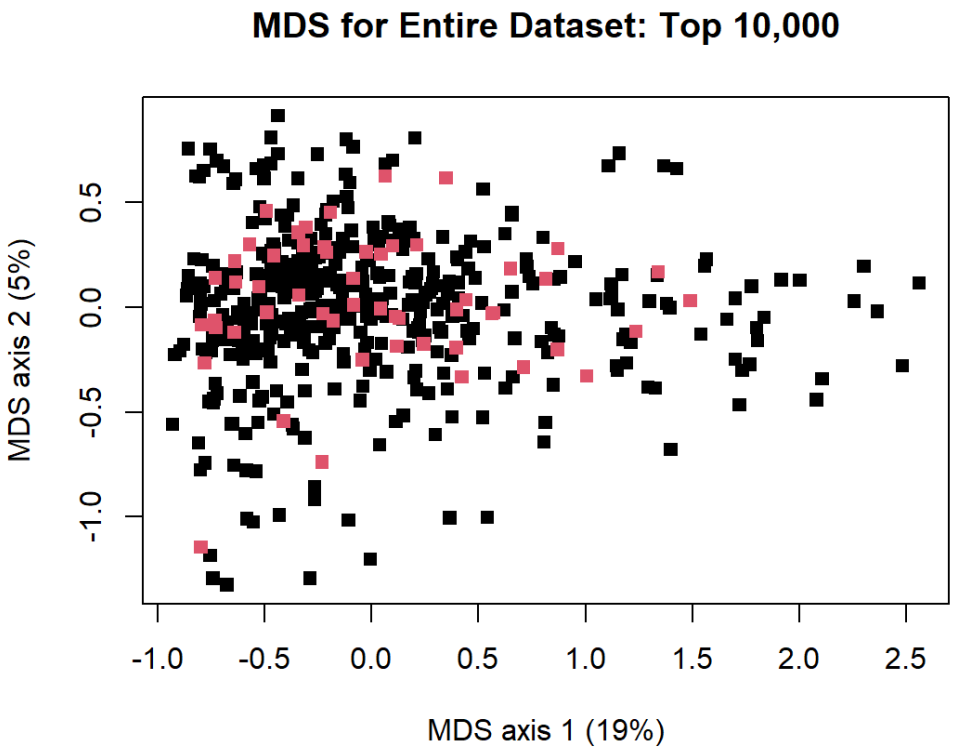

Figure 1. Multidimensional scaling plot of the top 10,000 common genes for the entire dataset. The two categories we can see are the positive (black) and negative (red) samples. There appears to be no clustering among the samples.

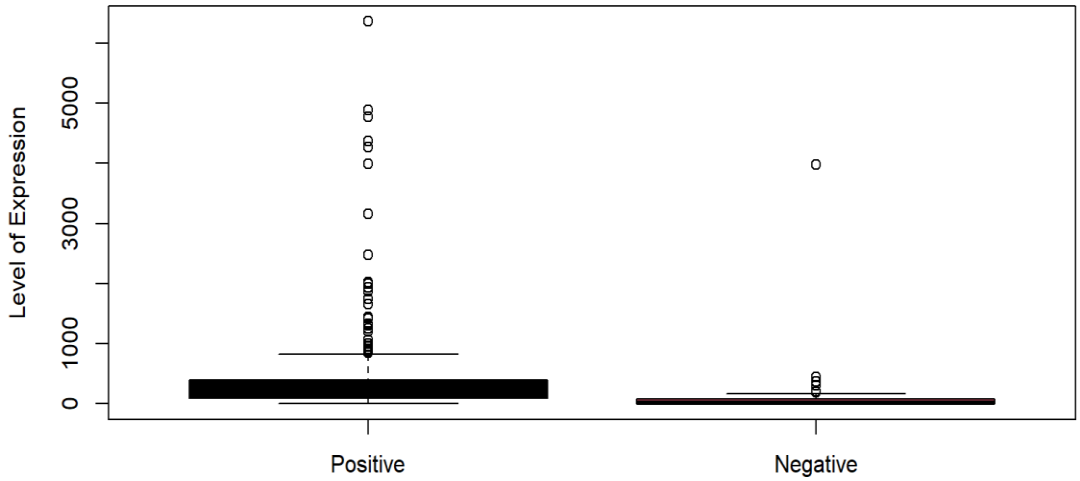

Figure 2. Level of Gene Expression in the XAF1 gene between individuals that test positive and negative for Covid-19. The DE gene is upregulated in the comparison between positive and negative individuals. This gene encodes a protein which binds to and counteracts the inhibitory effect of a member of the IAP (inhibitor of apoptosis) protein family.

Table 1. Table of the top differentially expressed genes that were upregulated and downregulated in the comparison between Covid positive and negative individuals. These are the genes that were used as features in the classifier.

UP		DOWN
XAF1	HERC6	AC110741.3
OAS2	GBP4	OR14L1P
OAS3	CXCL10	UBE2SP2
IFI44L	DDX58	IGKV1-8
IFIT1	CMPK2	SCGB3A1

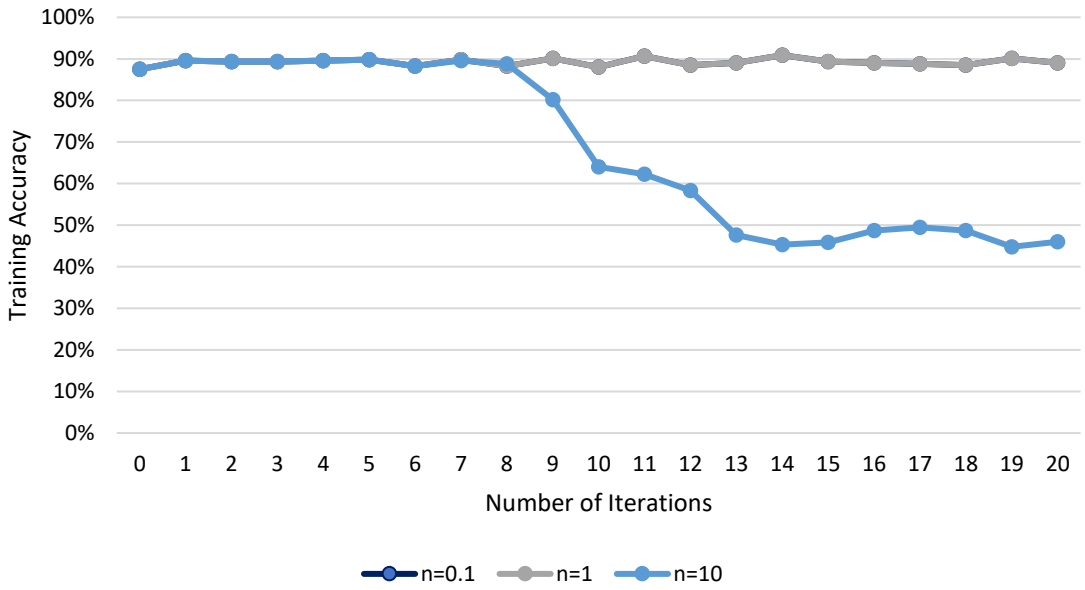

Figure 3. Training accuracies after each iteration for the learning rates n = 0.1, 1, and 10. Both n = 0.1 and n = 1 had the same training accuracy after each iteration. With the increase in learning rate, the training accuracy decreased 46% by the 20th iteration. The testing accuracies after using these learning rates for 20 iterations were 87%, 87%, and 27% respectively.

5. Discussion

When working with big data, I've realized that there are so many methods and analyses that can be performed on just a single dataset. There are many more things I would love to learn from this dataset, like looking at the top DE genes (Table 1) and finding their functions. Although I created a classifier that resulted in 87% accuracy (Figure 3), I would still want to see if I could incorporate the other variables from the data like viral load to see if I could improve the algorithm. I also learned that each dataset is unique including subsets of the original dataset. This has taught me that I should never settle for the default settings in my code but should test out all combinations to see which produces the most accurate/significant results. However, for next steps, I would love to actually have a dataset with even more patients because only have 484 samples for the classification algorithm which then had to be made even smaller into training and testing subsets. More data would help improve accuracy or help me discover new features of the data that were hard to find since they were present in such small quantities.

References

[1] Lieberman, N. A., Peddu, V., Xie, H., Shrestha, L., Huang, M.-L., Mears, M. C., Cajimat, M. N., Bente, D. A., Shi, P.-Y., Bovier, F., Roychoudhury, P., Jerome, K. R., Moscona, A., Porotto, M., & Greninger, A. L. (2020). In vivo antiviral host transcriptional response to SARS-COV-2 by viral load, sex, and age. PLOS Biology, 18(9). https://doi.org/10.1371/journal.pbio.3000849