Exploring Gene Expression Levels in Wound Healing

Alex Ruan
alexruan@umich.edu
Computer Science

Matthew Schneider
mrschnei@umich.edu
Computer Science

1 Introduction

For our final project, we wanted to see how inflammatory genes are differentially expressed (expressed differently) between normal mice and diabetic mice. We obtained our data from a study done at the University of Michigan where researchers collected loads of biological data on wounds from mice with and without diabetes.

In the original study the researchers conducted experiments on common house mice. They split the mice into two groups and fed on group a normal diet (ND) and fed the other group a high fat diet, to create diet included obesity (DIO). After creating the two groups, the researchers then created a wounded in each of the mice. Over the course of the week following the wound, they took DNA samples at two different stages, when they had high Ly6C expression and low Ly6C expression.

Ly6C is a monocyte/macrophage, which is a type of white blood cell that attacks anything that is not healthy to the body. Ly6C is a specific macrophage that is seen in wound healing and has two different stages of expression, Ly6C^{Hi} and Ly6C^{Lo}. Because this macrophage has different stages of expression it could potentially have different gene expression levels at those different stages.

The researchers also took duplicate samples for two groups of mice to account for variability, these groups will be referenced to as rep-1 and rep-2. As a result of the experiment, we now have four groups of data across two factors (DIO vs ND, Ly6C^{Hi} vs Ly6C^{Lo}) where each group is duplicated (rep-1 and rep-2).

The original goal of the research experiment performed was to explore how wound healing occurs in diabetic patients vs non-diabetic patients at the biological level. They believed this would be an important topic to research because it of its possible positive impact on society. This is helpful to society as hospitalization due to non-healing wounds because of chronic inflammation due to diabetes is one of the most common sources of hospitalization for those with diabetes, often resulting in amputation. With a 20-50% 5-year mortality rate post-amputation, it is very concerning that there is still no effective treatment for chronic inflammation due to diabetes. So, their goal from this study was to figure out what genes are associated with chronic inflammation in order to design treatments that targets those genes specifically, to treat chronic inflammation, and maybe improve treatment for those with diabetes.

When discussing the research study with the Dr. Aaron denDekker, who is a researcher in the University of Michigan Medical School, he told us that the researchers participating in the study did not do much deep analysis into the data and only focused on the surface level of the data from the experiment. This gave us the idea to explore the data deeply and to apply techniques that we learned over the course of the semester in EECS 476.

Our goal in exploring the data was to find what genes have the most difference in expression between DIO and ND mice. This idea led to us exploring what data points (genes) are the outliers in the distribution.

The idea of our project is that if we could find what genes are outliers, then we could then do a follow-up study to figure out why they are outliers, and if it relates to chronic inflammation at all. If this is not possible, then we can still identify genes that are outliers which could be helpful in future studies. Given this our findings could make it so that researchers are looking at the genes that relate more accurately to chronic inflammation for those with diabetes.

Over the course of working on our project we used multiple techniques that we learned in class. We used outlier, clustering, and dimensionality reduction techniques because we believed that they would be best for finding the results that we hypothesized. We attempted to use hierarchical clustering on the data to cluster it into groups, but we did not find this method effective, so we decided not to include it. For the one technique from the second half of the course, we used principal component analysis (PCA). For the two techniques that we had not already implemented in any of the three projects, we used PCA and extreme value analysis (EVA) for outlier detection.

2 Data

A majority of the data collection process was described in introduction section of this report. As a quick recap, the researchers created two different groups of mice, one group with a normal diet and one group with diet-induced obesity. They then created a wound on each of the mice and collected data each day for about a week on all of the mice to detect different expression levels of the macrophage Ly6C.

We received the raw form of this data, which was a total of 16.4 GB of RNA sequences given as fastq.gz files. In order to get a dataset that we could mine effectively, we ran through two different stages of preprocessing, the first was five steps that were given to us by Dr. denDekker.

The first step in preprocessing was to use FastQC, a quality control tool, to check reads for base quality. Then for any reads that have an adapter sequence or are poor quality, we used a trimming tool to remove any adapter sequence or poor quality reads because these RNA sequences are not what we want in our dataset. We then checked our quality again using FastQC to ensure our RNA sequences were good reads and clean. We then had to align the sequences to either a genome or a curated set of genes to identify what the exact sequences were. There are two popular alignment tools that are typically used in this field, Tophat2 and Hisat2. We used the Hisat2 program to align our raw RNA sequences because it was the aligner used in the original research paper. The final preprocessing step in this stage was to read the counts of each transcript. This gave us a dataset with a count for how many proteins produced by each gene were counted.

This dataset can be found in our Github folder under *results/count_output.txt.* Each row is a certain gene and includes the gene identification number and the reads for the 8 different cells that were tested (ND rep-1 Ly6C^{Lo}, ND rep-1 Ly6C^{Hi}, ND rep-2 Ly6C^{Lo}, ND rep-2 Ly6C^{Hi}, DIO rep-1 Ly6C^{Lo}, DIO rep-1 Ly6C^{Hi}, DIO rep-2 Ly6C^{Lo}, and DIO rep-2 Ly6C^{Hi}). Here ND is normal diet, DIO is diet induced obesity, Ly6C^{Lo} and Ly6C^{Hi} are two expression levels of the Ly6C macrophage, and rep-1 and rep-2 are just replicates (mice with the same conditions) to account for variation between mice. There are also some other columns in this dataset, but they were not relevant to our work.

We then performed our second stage of preprocessing where we got the data into a differentially expressed format. We took the dataset described above, dropped the irrelevant columns, and then calculated the log fold change between DIO-Ly6C^{Lo} and ND-Ly6C^{Lo}, as well as DIO-Ly6C^{Hi} and ND-Ly6C^{Hi}. An example of the formula that we used for both Ly6C^{Hi} and Ly6C^{Lo} is:

log(DIO_rep-1 + DIO_rep-2 + 1) - log(ND_rep-1 + ND_rep-2 + 1)

Using this formula, we obtained a value for the log fold change for both Ly6C^{Hi} and Ly6C^{Lo} macrophages. Our dataset that we used for our data mining then ended up have three columns. These columns were the gene identification number, the log fold change for Ly6C^{Lo}, and the log fold change for Ly6C^{Hi}.

3 Data Analysis

3.1 Initial Exploration

After doing all of our preprocessing steps, we then decided to explore the distribution of our data to help decide which mining methods would be most effective.

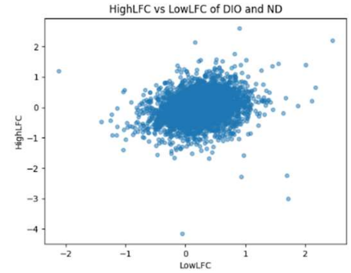

Figure 1: Scatter plot of log fold change between High and Low Ly6C

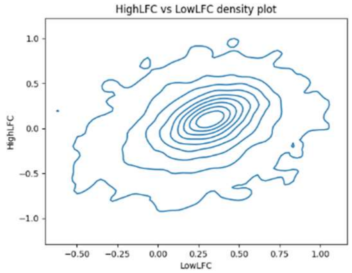

Figure 2: Density plot of log fold change between High and Low Ly6C, slightly zoomed.

Examining figures 1 and 2, we noticed two major things about the distribution of our data. In figure 1, we noticed that there seems to possibly be some outliers and that some methods of outlier detection might be effective for analyzing our data. In figure 2, we noticed that the data seems to follow a gaussian (normal) distribution. Using these ideas, we decided to first use extreme value analysis (EVA) on the data because our data seems to follow its main assumption, that the data is normally distributed.

3.2 Outlier Detection

As described in the last section, the first outlier detection method that we used was extreme value analysis. We decided to use this because of the distribution of the data. To do extreme value analysis, we found the mahalanobis distance for every point and for any point with a distance above three, we determined it to be an outlier. Figure 3 shows the scatterplot obtained from graphing the data in the same fashion as figure 1, but coloring in the points based on whether or not they are an outlier. As can be seen in the figure, we had a large number of outliers surrounding the central distribution of the data. In total we found 480 outliers.

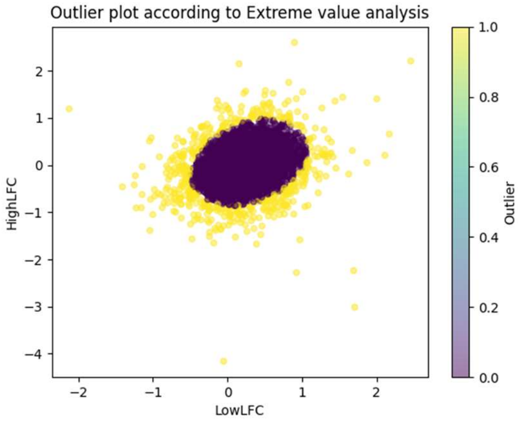

Figure 3: Outlier scatter using EVA. Yellow points are outliers.

The next outlier detection method that we used was Density-based spatial clustering of applications with noise (DBSCAN). We found this method from one of the slides in lecture six and decided to use it because it is a cluster-based method. After doing some research on how DBSCAN works, we found that DBSCAN automatically finds outliers in data by labeling any point not part of a cluster as an outlier. We used the Scikit-learn (sklearn) implementation of DBSCAN and trained our model to output the same number of outliers as the EVA outlier's method did. In order to do this, we tested different hyperparameters until we found one that gave us the correct number of outliers. We found that setting epsilon equal to about .0465 gave us the correct number of outliers that we were looking for.

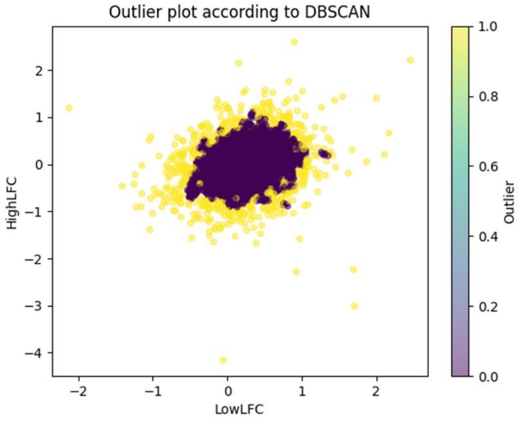

Figure 4: Outlier scatter using DBSCAN. Yellow points are outliers.

After finding the outliers using the two methods, we decided to find the Jaccard index between the two sets of outliers to see how similar they were. We got a Jaccard index of .686, showing that there is a good amount of similarity between the two sets of outliers, but still some differences.

While both plots are viable, we thought that the best approach would be to use extreme value analysis to find an outlier count, and then use that to tune DBSCAN to find that number of outliers. This is because the data is not perfectly smooth, and the outliers generated by extreme value analysis are too smooth. So, by using DBSCAN, we can account for the noisiness that is a part of the data.

3.3 Dimensionality Reduction

The final data mining technique that we used on our data was dimensionality reduction and more specifically PCA. Our goal with using PCA was to see how much variance was explained by each log fold change . We found that the variance explained by the log fold change of Ly6C^{Lo} and Ly6C^{Hi} were 0.6694 and 0.3305, respectively.

We then tried to find outliers using extreme value analysis using just the first component, log fold change of Ly6C^{Lo}, since it had the higher variance. We then determined any point above three standard deviations from the mean to be an outlier. This method only yielded 48 outliers, which is only 10% of what we found using the same method on the original dataset. Because of this we determined that while slightly more than ⅔ of the variation was explained by that one component, that one component by itself is insufficient to find outliers.

4 Conclusions & Discussion

Overall, we are happy with the outcome of our project. We were able to use a dataset from an official research study and run data analysis on it that the original researchers had not. Through our use of outlier detection methods and dimensionality reduction we found 480 genes that were outliers in their expression of Ly6C during wound healing in the mice. It would be very interesting to see how much meaning these genes have and if any of them could be used to identify a treatment that could help those with diabetes avoid having wound inflammation more often than others.

As a team, we faced a couple of challenges throughout the project. The biggest challenge that we faced was understanding the topic that we were focusing our project on. Neither of us are very knowledgeable about biology, and having to learn about the very complicated topics used in the original paper was difficult. Even now, we still have a basic understanding, but it was enough to be able to create a dataset and apply data mining techniques to it. We also faced challenges related to timing and communication. We were both very busy during this semester and finding time when both of us were available was difficult.

Doing this project, we learned a lot about biology and bioinformatics. It was difficult at times to wrap our heads around the topics, but we succeeded at gathering a basic perspective of the data. We also used data mining techniques that we had never before used, like DBSCAN and EVA. These outlier detection methods turned out to be very interesting and useful for what we wanted to do with our project. Using these methods that we had not used before was also our most liked part of the project. Neither of us had used these techniques in actual practice, and by using them and seeing how they determined outliers graphically was very interesting.

For our project we tried to split it as evenly as possible. Alex was the one who came up with the idea to get a dataset from the bioinformatics lab and setup the meeting with Dr. denDekker. After we received the data and the ideas from Dr. denDekker, we tried our best to communicate effectively with each other and brainstorm which data mining techniques would be most effective for analyzing the data. After brainstorming ideas, Alex did most of the coding for the techniques that were described in the earlier sections, while Matt wrote a majority of the final report.

5 References

[1] Kimball, A., Schaller, M., Joshi, A., Davis, F. M., denDekker, A., Boniakowski, A., Bermick, J., Obi, A., Moore, B., Henke, P. K., Kunkel, S. L., & Gallagher, K. A. (2018). Ly6CHi Blood Monocyte/Macrophage Drive Chronic Inflammation and Impair Wound Healing in Diabetes Mellitus. Arteriosclerosis, thrombosis, and vascular biology, 38(5), 1102–1114. https://doi.org/10.1161/ATVBAHA.118.310703