Gene Expression Analysis in THOR and Methods

4.28.2023

—

Mentor: Prof. Julia Fukuyama

Benjamin Shores

Morgan Fissel

Introduction

Project Introduction

In this project, we will be analyzing an existing dataset to attempt to reach the same conclusions drawn by previous analysis of the dataset to solve the analytical problem of how three bacteria influence their gene expression in a microbiome.

Background and Importance

The paper “THOR’s Hammer: the Antibiotic Koreenceine Drives Gene Expression in a Model Microbial Community” analyzes gene expression of The Hitchhikers of the Rhizosphere, or “THOR”, an important part of the root ecosystem in some agriculturally important plants such as soybeans. Recreating these results with our own methods will help to verify results already found and determine relative accuracy of methods used to analyze gene expression.

Introduction to Domain

- A microbiome is the contained microorganisms and their interactions within a particular system.
- Not all genes are used the same amount, depending on a cell’s needs and its context. The concept of how much certain genes are being used, or “expressed”, is measured through volume of RNA, a single-stranded macromolecule that is the “messenger” between DNA and the cell.
- The process of going from a collection of bacteria to numbers representing the RNA found in them is complex and goes through many processes, though widely considered accurate; it is not as simple as counting, so RNA sequencing data necessarily comes heavily cleaned.

Assessment

As we are working to reproduce results from a paper, that paper may constitute the current solution.

We assume that data is fit to the model, though this may not necessarily be true. We also assume that the original paper draws largely correct conclusions as a baseline. Our available resources appear to be adequate for analysis. Our analysis of models will be constrained by the fact that we are analyzing off of only the THOR dataset.

Research Questions

- How do the different members of the microbial communities affect each other?
- Which genes are most often expressed together?
- Do the different models represent the communities accurately?

Business Objectives and Success Criteria

- Find groups of genes in an important trio of bacteria which are affected by the presence of other members of the trio. A report of which factors influence which genes in which ways should constitute success for this objective.
- Create LDA, MIMIX, and possibly other models for the data. Compare said models to the previous study. Success for this objective will be achieved when we can discuss the ways these methods were employed and the pros/cons of each.

Data Mining Objective and Success Criteria

- Acquire familiarity with gene expression analysis methods. Make the models as accurate as possible. Make the models legible. We will consider this objective achieved when we have created models for at least LDA and MIMIX which can be compared to the original study

Scope and Limitations

The scope of the project is contained within the amount of different models we would be making. We currently have two, with potential for more if needed.

Project Plan

1. Choose and familiarize with a method
2. Apply method to the data set
3. Analyze results
4. Compare to previous results
5. Repeat for next method(s)
6. Draw conclusions

Data

Description

Data was collected from roots of field-grown plants. Gene expression data was then derived using sequencing methods which, as described earlier, makes the data heavily cleaned.

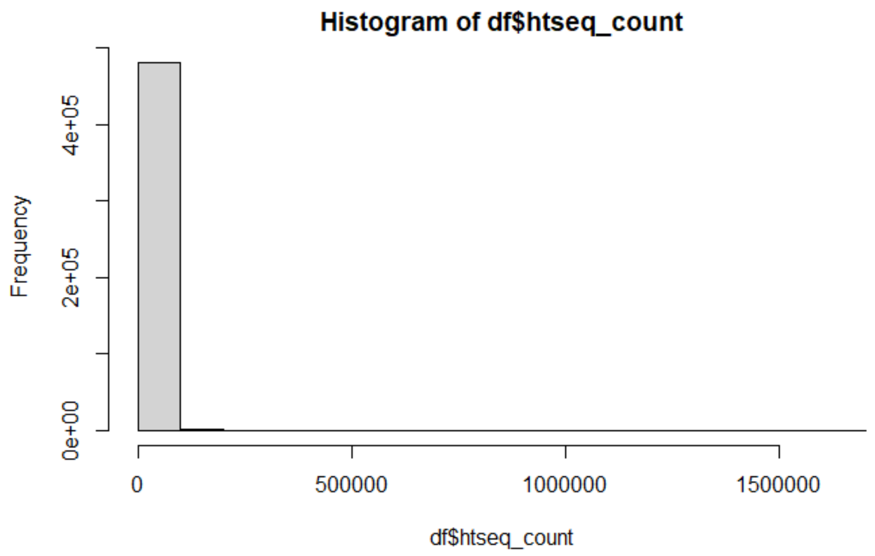
Most of the data follows the format given below, with each line representing data on a given gene for a given sample. There are two sets, one which represents the gene expression found in the THOR microbiome as normal with 482832 instances and the 5 attributes given below, and another which represents the gene expression found in a version of the microbiome with certain genes suppressed, with 620784 instances and the 5 attributes given below.

Attribute Name	Type	Description
Sample	Categorical	Indicator of the sample from which the line of data was collected
Condition	Categorical	Environment the sample was taken from (which other bacteria were present)
Organism	Categorical	Name of the organism the sample was taken from
Gene	Categorical	Name of the gene measured
htseq_count	Quantitative	Quantity of mapped reads per gene

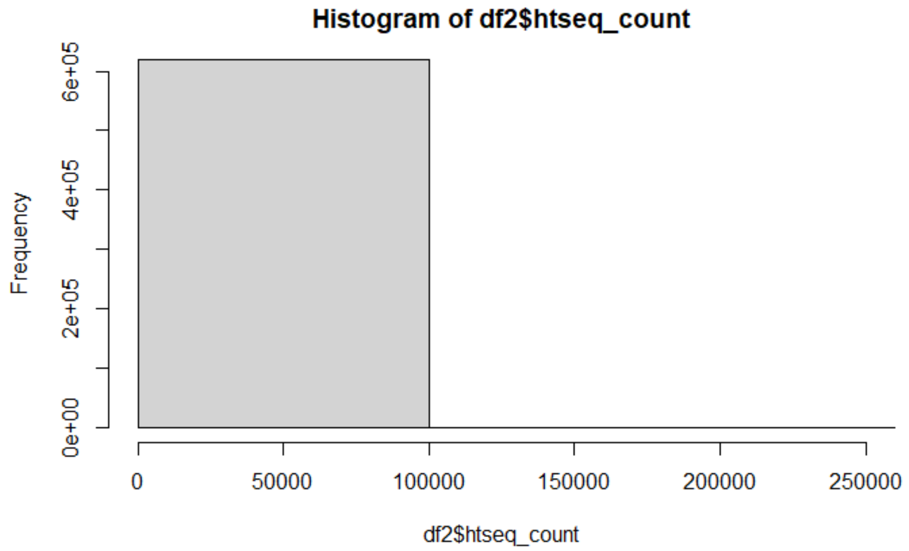
We were also provided with a dataset which gives known groups of orthologous genes. It has 17244 instances and 2 attributes given below.

Attribute Name	Type	Description
Gene	Categorical	Name of the gene
Cog_Category	Categorical	Gene COG (Cluster of Orthologous Gene)

Exploration



Histogram of the number of reads from the unaltered THOR microbiome

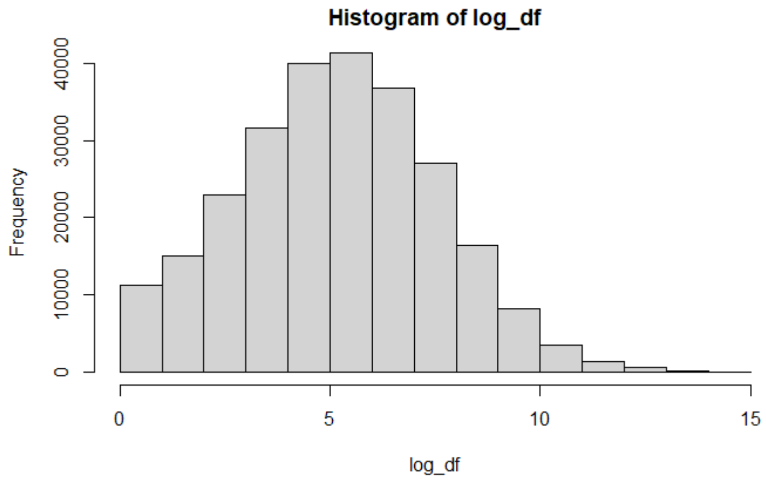


Histogram of the number of reads from the altered THOR microbiome

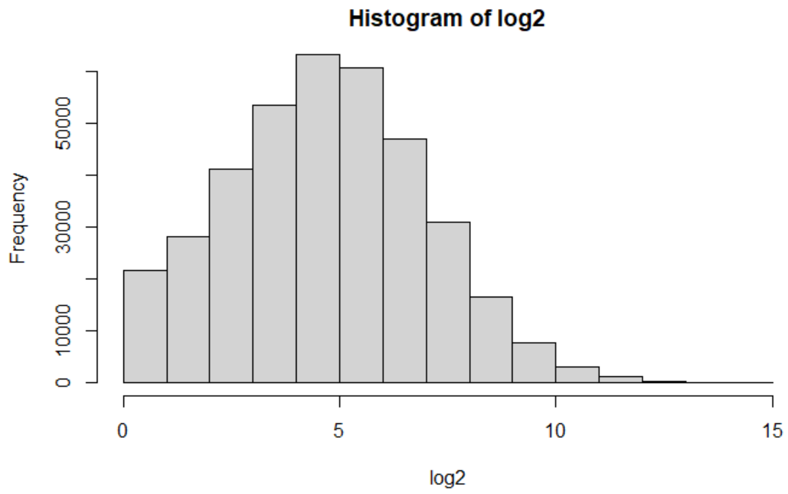
As illustrated by the above graphs, the data has a pretty nasty right skew that makes basic statistical descriptions of the data nonsensical. In fact, 40% of genes in the unaltered microbiome and 47% of genes in the altered microbiome have counts of 0. As such, for this portion of the project, we applied a log transformation for the purposes of deriving basic statistical descriptors of the data, which you can see below.

Dataset	Min	Q2	Median	Mean	Q3	Max
Altered Microbiome	-Infinity	-Infinity	1.386	-Infinity	5.347	14.304
Unaltered Microbiome	-Infinity	-Infinity	2.398	-Infinity	5.193	14.214

Keep in mind that these are the results after a log transformation. A value of -Infinity represents 0. All in all, even a log transformation is not enough to mitigate the influence of that nasty skew. From this high level view, it does not seem like there is much difference between the two datasets, though the differences in mean could represent something significant.



Histogram of the log of the number of reads from the unaltered THOR microbiome



Histogram of the log of the number of reads from the altered THOR microbiome

These histograms of the log-transformed data support the conclusions drawn above, Though the log transformation does work to normalize the distribution, the effects of the heavy skew are still apparent. The distribution between the two data sets appear largely the same.

Preparation

- The data has significant outliers and a heavy skew, but these are expected in gene expression data. The large numbers of zeroes are also to be expected in gene expression data. This is because many genes in a genome will not be expressed at any given time.
- Because of the process of finding gene expression data from a microbiome, the data is already heavily cleaned.
- A log transformation was employed for the graphs for illustrative purposes, but we do not plan on applying this transformation.
- For its analysis, the paper also considered all genes with counts of less than 10 to be zero to make their analysis less complex. We do not currently plan on doing this but it may nevertheless come up.

Model Building

LDA Model 1

Our first LDA model fit using R was not necessarily built to give intelligible results. It was built during a period of the project that would be best characterized by the many setbacks trying to get the proverbial wheels rolling and was needed to verify that the LDA could be run at all. Default parameters were used and the number of topics was chosen as 5 arbitrarily. Topics were found but not analyzed for this reason.

LDA Model 2

The second LDA model fit using R had parameters given as the result of LDA tuning methods. After learning how to implement and run these methods, we got a model that we feel relatively confident about. Based on the results of the LDA tuning methods, we set our number of topics at 4 for both data sets, though a version with 5 was also found to possibly be worth investigating for the htseq_delkec data set. From this we derived topics most frequently associated with the presence of certain microbes in the THOR microbiome.

MIMIX

Our next model to investigate was MIMIX, a Bayesian mixed-effects model made for microbiome data analysis which treats the data as a response variable to a designed experiment. The original MIMIX code found here https://github.com/nsgrantham/mimix had depreciated and so our mentor provided a fixed version which may be found here https://github.com/jfukuyama/mimix. Except for the number of iterations, which was cut in half from 20000 to 10000 as a concession to time constraint and which we are assured is still more than enough to get accurate results, parameters were kept at the default used in the test case. Once it was running, we found no need to iterate further on the model.

Results

LDA

As it was not made to be a source of conclusions, the first LDA model was not taken as a source of any conclusions. From the second model, however, we were able to derive conclusions.

The results of the htseq_wt data were very clean. We found 4 topics of genes, 3 of which we suspect belong generally to the three members of the microbiome and the fourth being the genes which we believe are expressed in interaction between members of the community. Behavior in the grouping among these genes that were associated with individual members of the microbiome was also able to be explained with the frequency of reads mapped from those members. All of these conclusions showed strong agreement with the original paper's conclusions.

The results from the htseq_delkec data were a little sloppier, perhaps because the data was not so conducive to this method of analysis due to the low-inoculum population. With more time, refining to get cleaner results for this dataset probably would have been the first priority. That said, we found a topic which we have reason to believe may represent genes expressed due to interaction after the gene suppression, a phenomenon the original paper notes- essentially, gene expression still changed with interaction, but in the opposite direction.

MIMIX

MIMIX results are much more straightforward. Though the MIMIX method gives a lot of output that may be used for analysis, we are currently looking at a single boolean variable in the output: reject_global_null. This variable tells whether the evidence is enough to reject the null hypothesis that there are no connections in the groups. If this variable is true, then we can conclude the alternative hypothesis that there are connections in the groups.

For htseq_wt, this variable was true, which likely points to the particular interaction noticed through LDA grouping. This is consistent with the finding of the original paper.

Running htseq_delkec with MIMIX also returns a true variable, which would seem to disprove the idea that the suppressed proteins cause the interaction in the THOR microbiome, but we return to the findings of the original paper that interaction with altered populations changed gene expression but in a different way. It is likely that this result is from MIMIX finding that interaction.

Comparison

Both models seemed to draw conclusions consistent with previous analysis of the data and each other. What may then distinguish them? MIMIX was more automated and therefore convenient once it was able to be run. However, from the vantage point of analysis done during this project, it would seem that LDA gave greater insight into the data. While the htseq_wt dataset was straightforward, the htseq_delkec dataset was messier to analyze as its contrast and the reason was only clear through LDA: there was still interaction with the gene suppressed but it was of a different kind. MIMIX, in drawing a cleaner boolean conclusion, made more of the reason for the conclusion lost in the mathematical process. For this reason alone we may say LDA was better suited to the task at hand.

It ought to be said of MIMIX that this conclusion may be unfair. Despite the above sections drawing conclusions only from the reject_global_null variable, it is far from the only output of MIMIX's analysis. Local-estimates, for example, are given and could be used to look into the most upregulated genes in a certain group. Professor Fukuyama also points out that it gives coefficient estimates that could be compared to the topics that LDA found. One with more proficiency in reading its output or with more time to learn how to interpret it may very well extract much greater insight and be able to say that MIMIX was more suited to the task.

Evaluation

We have employed both MIMIX and LDA in models to find in gene expression data the interaction between members of the focus microbiome. Both models meet success under the business objective. Since both models draw conclusions consistent with previous analysis of the data, we can also say that both meet the data mining objective.

Conclusion and Future Work

Throughout this project we familiarize ourselves with tools, implementations, and uses of MIMIX and LDA as applied in the realm of data science in bioinformatics, finding meaning in a mass of biological data too vast for traditional human analysis. Both methods gave conclusions consistent with the original findings of the first paper on the data and pointed to the clustering of gene expression data such that it became clear that there were a group of genes regulating interaction within the microbiome from which the data was gathered.

This work opens up many areas for further work. First, we were open to the possibility of comparing to a third or fourth kind of method if more time had been available. This is work that may still be pursued. The LDA for the htseq_delkec data may also have been explored for a better fit. With even more time, one might begin looking into MIMIX and/or LDA results to begin looking into the specific genes and what, if any, known uses they have, tying it back into more general biological knowledge. Lastly, it is within possibility to say that an entire project could be done on MIMIX, both in familiarizing oneself with the conclusions to be derived from its output and in improving its usability.

Bibliography

Grantham, N. S., Reich, B. J., Borer, E. T., & Gross, K. (2017, March 22). Mimix: A bayesian mixed-effects model for microbiome data from designed experiments. arXiv.org. Retrieved December 6, 2022, from https://doi.org/10.48550/arXiv.1703.07747

Hurley, A., Chevrette, M. G., Rosario-Melendez, N., & Handelsman, J. (2022, April 18). THOR's Hammer: the Antibiotic Koreenceine Drives Gene Expression in a Model Microbial Community. American Society For Microbiology. Retrieved December 6, 2022, from https://journals.asm.org/doi/full/10.1128/mbio.02486-21

Kris Sankaran, Susan P Holmes, Latent variable modeling for the microbiome, Biostatistics, Volume 20, Issue 4, October 2019, Pages 599–614, https://doi.org/10.1093/biostatistics/kxy018

Liu, L., Tang, L., Dong, W. et al. An overview of topic modeling and its current applications in bioinformatics. SpringerPlus 5, 1608 (2016). https://doi.org/10.1186/s40064-016-3252-8