IBD Final Paper

Ayal Meir Yakobe and Harrison Fried

May 2024

1 Abstract

Inflammatory Bowel Disease (IBD) is a multi-factorial disease involving genetic, environmental, and microbial factors that results in inflammation of the gastrointestinal tract. Of particular interest in recent years has been the relation between IBD and the microbiome, with evidence signaling significant changes in taxa abundance and microbial diversity within the intestinal flora of IBD patients. To study this phenomena, 16S sequencing data was extracted from the NCBI database, analyzed using DADA2 to tabulate taxa and relative abundance data, and then fed into Random Forest Classifier to check whether an individual's microbiome data can be predictive of IBD. Lastly, efforts were made in order to allow for prediction of the future microbiome state of an individual based on time-series data.

2 Introduction

Inflammatory Bowel Disease (IBD) represents a group of autoimmune diseases characterized by chronic inflammation of the gastrointestinal tract, affecting millions worldwide. The pathogenesis of IBD, encompassing Crohn's disease and Ulcerative Colitis, is complex and influenced by genetic, environmental triggers, and factors related to the microbiome. Early and accurate prediction of IBD can significantly enhance patient management, prevent severe complications, and improve quality of life. However, traditional diagnostic methods are often invasive and costly, necessitating the exploration of alternative predictive approaches. [McDowell et al., 2023] [Lloyd-Price et al., 2019]

Recent advancements in data science and bioinformatics have opened new avenues for disease prediction and management. The National Center for Biotechnology Information (NCBI) hosts a vast repository of genomic and microbiome data, which provides an unprecedented opportunity to harness big data for predictive analytics in IBD. By leveraging this rich dataset, researchers can identify biomarkers and construct predictive models that discern subtle patterns indicative of the disease. [ncb, 2024]

This study aims to develop a predictive model for IBD using machine learning techniques on a large-scale dataset from NCBI. The model intends to predict the likelihood of a patient having IBD by analyzing the relative abundance of specific microbial features extracted from 16S sequencing data. We hypothesize that the integration of comprehensive bioinformatics resources and advanced machine learning algorithms will allow for a possibly non-invasive, accurate, and early diagnosis of IBD, thereby facilitating timely and personalized therapeutic interventions.

In the following sections, we will outline the data acquisition process from NCBI, describe the methodology for feature selection and model building, present our findings, and discuss the implications of our results in the broader context of gastroenterology and personalized medicine.

3 Methodology

3.1 Literature Review

Literature review on IBD and on the microbiome's relation to IBD was performed. Following literature review, studies on the IBD microbiome with publicly accessible 16S sequencing data were found [Abdel-Rahman and Morgan, 2023]. These studies were further filtered down based on whether the sequencing technique was compatible with DADA2 and had relevant metadata available. From studies with usable data, we collected BioProject links from the National Center for Biotechnology Information (NCBI) database, extracting relevant information for our analysis [ncb, 2024]. Subsequently, we compiled this information into a structured format using Google Sheets, allowing for easy organization and manipulation of the data.

3.2 Data Extraction

The microbiome data utilized in the script, fastq files, were extracted from the National Center for Biotechnology Information (NCBI). NCBI plays a pivotal role in advancing science and health by making biomedical and genomic information accessible. A python script was utilized to scrape SRA accession numbers for relevant BioProjects from the NCBI database. Following this, a bash script extracts the specified datasets related to microbiome studies using fasterq-dump to download .fastq files, which are then processed and analyzed to derive insights into the underlying biological and health-related questions. For some studies, further purification of accession lists had to be performed to get only those samples that had metadata and were the appropriate type of sequencing method. [ncb, 2024] [Lloyd-Price et al., 2019] [Liu et al., 2016] [Gevers et al., 2014]

3.3 Abundance Calculation Using DADA2

DADA2 was chosen for abundance calculation for its increased sensitivity to rare taxa within samples. DADA2 works by taking fastq files as input with paired ends separated into different files. These ends are then separately filtered, trimmed, error rates were learned, and de-replicated. Following de-replication, DADA2 was performed to get the abundances of these unique sequences, and forward and reverse ends were merged. Further filtering was performed to get a final amplicon sequence variant (ASV) table, which were then assigned taxa based on SILVA V138.1 reference database and imported into phyloseq objects [Quast et al., 2013]. DADA2 was performed separately for each study due to different read lengths and the expectation of different error rates. When merging studies, phyloseq objects were merged in order to get final merged abundance tables. For the ”Gevers 1” study, due to a very large number of samples, the de-replication process crashed due to exceeding RAM capabilities, so samples were batched into 6 batches of 150 samples each and then merged as phyloseq objects. [Prodan et al., 2020] [Callahan, n.d.]

3.4 Pre-Processing Abundance Data

Following DADA2 abundance calculations, pre-processing of the data was required to attach metadata to samples, remove samples with 0 abundance found, and normalize to achieve relative abundance for each sample. Publicly accessible metadata for each study was collected from NCBI BioProjects as well for some from supplementary files for the study. The ”Gevers 1” study required metadata accessed from Qita as well. For each sample using a Python script, subject ID, collection week, IBD status, sex, age, where study the sample is from, and the type of sample (stool, biopsy, etc.) were attached to the sample abundance data. For studies lacking time-series data, collection week was set to 0 for all samples. Next, a few samples that lacked any abundance data were removed. Following this step, row sums were calculated and rows were divided by row sums to obtain relative abundance data and normalize across samples. [Lloyd-Price et al., 2019] [Liu et al., 2016] [Gevers et al., 2014]

3.5 Classification

We chose the *Random Forest* classifier, a powerful ensemble learning method that combines the predictions of multiple decision trees. Generating a host of decision trees based on various features, all are then subsequently trained on a random subset of the data, collectively voting on the final prediction. This aggregation of decision trees enables the *Random Forest* classifier to handle complex relationships in the data and make accurate predictions with minimal computational overhead. Furthermore, the *Random Forest* classifier allows for a model that accounts for overfitting, inconsistent and sparse data in rows, and calculates feature importance. [ibm, 2024]

3.6 Prediction of Future Microbiome State (add)

In order to attempt to be able to predict the future state of the microbiome, Vector Autoregression (VAR) was implemented, a statisitcal modeling technique designed to capture the linear interdependencies among multiple time-series datasets [MathWorks, 2024]. This model would ideally uncover relationships between varying bacterial abundances over time.

Further, In order to reduce the dimensionality of the data, the 50 most important taxa from the Liu/Lloyd-Price/Gevers dataset were extracted and subjects with time series data were input into the model.

4 Results

Usable accessions were downloaded via the above methods from Liu et al., Lloyd-Price et al., and the first BioProject of Gevers et al. from the NCBI database. Following filtering and DADA2, 1168 samples were available for pre-processing. Pruning samples with total abundances of 0 resulted in 1164 samples remaining. Of these, 82 samples are from Liu et al., 176 samples are from Lloyd-Price et al., and 906 samples are from the first BioProject of Gevers et al. Following calculation of relative abundances and attachment of metadata to samples, the data was fed into *Random Forest* classifier. When only the Liu and Lloyd-Price data were input, the accuracy of the model was .81, with an average cross-validated accuracy of .62 signaling sensitivity to certain subsets of the data. However, upon including the Gevers data, the accuracy dropped to .64 with an average cross-validated accuracy of .66. When the Gevers data was input alone the accuracy was .71 with an average cross-validated accuracy of .67. For the model on the Liu and Lloyd-Price data, of the 20 most important taxa for prediction, 18 belong to the five most variable families of bacteria in abundance between IBD and control according to literature. For each model, the number of the top 20 most important taxa for prediction that are from one of the 5 most variable families in abundance between IBD and control was calculated according to literature [Alam et al., 2020]. For the Liu and Lloyd model, 19 of the top 20 most important taxa were from these 5 families. For the combined Liu, Lloyd and Gevers model, 17 of the top 20 most important taxa were from these 5 families. Lastly, the Gevers data alone model similarly had 17 of the top 20 most important taxa from these 5 families. Interestingly, there was significant variation in the distribution of these among 5 families between the datasets with number of important taxa from the Lachnospiraceae family decreasing and from the Bacteroidaceae family increasing from the Liu/Lloyd-Price dataset to the Gevers only dataset. [Lloyd-Price et al., 2019] [Liu et al., 2016] [Gevers et al., 2014]

5 Discussion

The increasing availability of sequencing data combined with the increasing ability of computational tools has allowed for computational methods to be a burgeoning tool for medical evaluation, classification, and prediction. In this study, these tools were applied to IBD, specifically 16S sequencing data of the intestinal microbiome from stool and biopsy samples of IBD and control subjects.

5.1 Conclusion

Following literature, a large number of candidate studies were identified. However, further screening significantly reduced the number of viable studies for our methodology, however still leaving thousands of samples. Additionally, due to computational limitations and time limitations, only a portion of these samples were able to be extracted from the NCBI database as downloading a single study's worth of data took up to days worth of time to complete. For this study, 1164 samples passed all steps, split between three different studies.

Following inputting various subsets of the data into the *Random Forest* classifier, these models were evaluated. The Liu/Lloyd-Price model has a relatively high accuracy of .81, although it appears to be suffering from over-fitting based on the fact that the average cross-validated accuracy is significantly lower. Upon significantly increasing the sample size by adding in Gevers data, the accuracy dropped significantly. This may be due to many different reasons. It may be a symptom of the over-fitting of the original model, an effect of the batching required to perform de-replication without exceeding RAM limitations, true differences between studies, or lower quality reads and/or poor quality sampling. Upon investigating the Gevers data alone model, while the accuracy was lower than the Liu/Lloyd-Price model, the average cross-validated accuracy was higher than that of the Liu/Lloyd-Price model. This points to differences between studies clouding accuracy rather than issues with the Gevers data. Additionally, accuracy differences between the models is likely not due to the ratio of IBD to control as the the change is within 1-2%.

In order to evaluate the performance of the classifier appropriately identifying critical taxa, the 20 most important taxa features of the models were compared to the 5 most variable in abundance taxa families, according to literature, between IBD and control patients [Alam et al., 2020]. Overall, all models tended to identify the most important taxa as part of these main 5 families, ranging from 19/20 to 17/20, with the Liu-Lloyd-Price model having the most. While taxa outside of these major 5 families may be truly important for prediction, for evaluating the model to ground truth, it can be reasonably expected that the families that change the most between IBD and control should at least be over-represented in the most important taxa, although it may not be a good evaluator between models both identifying a disproportionately high number of taxa as part of these families. Overall, the Lachnospiraceae family appears to be the most important taxonomic family for classification. Lachnospiraceae and Ruminococcaceae, one of the other main 5 families, are both involved in short-chain fatty acid (SCFA) metabolism and production, and have been shown to act as a chemical messenger to effect Treg differentiation, which is an integral part of IBD pathogenesis via inbalance in the T_{reg}-T_H17 cell axis in the intestinal epithelium [Vacca et al., 2020].

5.2 Future Work

In this study, we successfully developed a classifier model to predict inflammatory bowel disease (IBD) cases using extensive microbiome data. However, our efforts to construct a more complex model predicting future microbiome states faced challenges. We attempted to employ a VAR model but did not achieve meanigful results.

Despite the promising approach, we struggled with limited time-series data. After pre-processing to focus on the 50 most abundant bacteria and removing rows with missing values, we had 903 data points across 53 columns. While sufficient for a single subject, our dataset comprised only 202 repeating subjects with an average of 2.8 observations each for those that did repeat, however some of these are at the same time point. This limitation hindered our ability to apply VAR effectively. Future work will require more extensive time-series data, and additional resources in regards to time, memory, and computational capacity. Furthermore, due to differences between studies which were difficult to overcome, a unified study with vast amounts 16S sequencing samples would likely help classification accuracy. Additionally, incorporation of multi-omics data may help the accuracy of the model due to the multi-factorial nature of IBD.

References

What is random forest?, 2024. URL https://www.ibm.com/topics/random-forest. Accessed: 28 April 2024.

Welcome to ncbi, 2024. URL https://www.ncbi.nlm.nih.gov. Accessed: 28 April 2024.

Lama Izzat Hasan Abdel-Rahman and Xochitl C Morgan. Searching for a consensus among inflammatory bowel disease studies: A systematic meta-analysis. *Inflammatory Bowel Diseases*, 29(1):125–139, 2023. doi: 10.1093/ibd/izac194.

M.T. Alam, G.C.A. Amos, A.R.J. Murphy, et al. Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels, 2020. URL https://doi.org/10.1186/s13099-019-0341-6. Accessed: 5 May 2024.

Benjamin J. Callahan. Dada2 tutorial: Dada2 pipeline tutorial (1.8). https://benjjneb.github.io/dada2/tutorial$_{18}.html, n.d. Accessed 5 May 2024.$

Dirk Gevers, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, Xochitl C. Morgan, Aleksandar D. Kostic, Chengwei Luo, Antonio González, Daniel McDonald, Yael Haberman, Thomas Walters, Susan Baker, Joel Rosh, Michael Stephens, Melvin Heyman, James Markowitz, Robert Baldassano, Anne Griffiths, Francisco Sylvester, David Mack, Sandra Kim, Wallace Crandall, Jeffrey Hyams, Curtis Huttenhower, Rob Knight, and Ramnik J. Xavier. The treatment-naive microbiome in new-onset crohn's disease. *Cell Host & Microbe*, 15(3): 382–392, 2014. ISSN 1931-3128. doi: 10.1016/j.chom.2014.02.005. URL `https://www.sciencedirect.com/science/article/pii/S1931312814000638`.

Ta-Chiang Liu et al. Paneth cell defects in crohn's disease patients promote dysbiosis. *JCI Insight*, 1(8):e86907, June 2 2016. doi: 10.1172/jci.insight.86907.

Jason Lloyd-Price et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, 2019. doi: 10.1038/s41586-019-1237-9.

MathWorks. Introduction to vector autoregressive (var) models, 2024. URL `https://www.mathworks.com/help/econ/introduction-to-vector-autoregressive-var-models.html`. Accessed 5 May 2024.

C. McDowell, U. Farooq, and M. Haseeb. Inflammatory bowel disease. `https://www.ncbi.nlm.nih.gov/books/NBK470312/`, August 2023. Updated: August 4, 2023. Accessed: May 6, 2024.

Andrei Prodan, Valentina Tremaroli, Hanna Brolin, Aeilko H. Zwinderman, Max Nieuwdorp, and Ed Levin. Comparing bioinformatic pipelines for microbial 16s rrna amplicon sequencing. *PLoS One*, 15(1):e0227434, Jan 16 2020. doi: 10.1371/journal.pone.0227434.

Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 2013. doi: 10.1093/nar/gks1219.

M Vacca, G Celano, FM Calabrese, P Portincasa, M Gobbetti, and M De Angelis. The controversial role of human gut lachnospiraceae. *Microorganisms*, 8(4):573, Apr 2020. doi: 10.3390/microorganisms8040573.