An automatic machine learning framework for the analysis of microbiome data and robust pipeline identification and evaluation

By

Michael Greer

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

August 31, 2020

Nashville, Tennessee

Approved By:

Bing Zhang, Ph.D.

Danyvid Olivares-Villagomez, Ph.D.

Qi Liu, Ph.D.

Jacob Hughey, Ph.D.

James Crowe, MD

Cindy Gadd, Ph.D., MBA, FACMI

DEDICATION

To my parents and siblings

ACKNOWLEDGEMENTS

Many people have made significant contributions towards the completion of this academic milestone. First and foremost, I would like to thank Dr. Bing Zhang for his patience, encouragement, and mentorship. He has been an amazing mentor over the years and I would not have been able to complete this journey without his help. He has influenced my life in more ways than one and I am grateful for the time and energy he has put into ensuring I succeed.

I would also like to thank Dr. Danyvid Olivares-Villagomez, who acted as a co-mentor and introduced me to bench science. His scientific enthusiasm has been contagious and he has been an excellent co-mentor over the few years we've been able to work together. I am grateful he allowed me to be a part of his lab and learn what it's like to combine computational and bench science. Working in that setting was an incredible learning experience and I will always be appreciative for that opportunity.

I would like to thank the National Library of Medicine for financially supporting this work through their T15 training grant and also Vanderbilt's Department of Biomedical Informatics for providing a world-class learning environment. In particular, I would like thank Dr. Cindy Gadd, Rischelle Jenkins, and Dr. Kevin Johnson. Dr. Gadd has always been there to discuss life updates and offer advice and for that I am grateful. Rischelle Jenkins has been there to guide me through the various administrative tasks associated with being a student which has always been appreciated. And lastly, I would like to thank Dr. Kevin Johnson for his encouragement and guidance.

I would also like to express my sincere gratitude to my other committee members: Dr. Qi Liu, Dr. Jake Hughey, and Dr. James Crowe for their insightful comments and suggestions on ways to improve my research. This group of mentors have been incredibly supportive during these unique times and each of them have contributed to my work in their own way.

Finally, I would like to thank my parents and siblings. They are the reason behind most of my success and without them I wouldn't have made it this far.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUTOML - Automated machine learning

DNA – Deoxyribonucleic Acid

FS – Feature selection

GSEA – Gene Set Enrichment Analysis

IEL – Intra Epithelial Lymphocytes

KS – Kolmogorov Smirnov

ML – Machine Learning

NGS – Next generation sequencing

NLM – National Library of Medicine

PCA – Principle Components Analysis

QC – Quality Control

RNA – Ribonucleic Acid

AUC – Area Under Receiver Operating Characteristic

SPP-1 – Osteopontin

CHAPTER I

INTRODUCTION

The promise of machine learning and clinical bioinformatics

Clinical bioinformatics has the potential to transform healthcare by providing patients with personalized diagnostic and therapeutic solutions (Ashley et al., 2016, Hamburg et al., 2010). The pace of innovation within this field has been staggering with consistent improvements in both sequencing technology and analysis methodology. For example, we are currently in the "fourth generation" of next generation sequencing (NGS) technology and current platforms are now capable of generating long multi-kilobase reads which was previously impossible (Slatko et al., 2018). Additionally, new methods of analyzing data are constantly being published which help drive both basic science research and diagnostics (Kelly et al., 2016, Sajda et al., 2006, Zhou et al., 2015). Given these advancements, it is not unreasonable to expect a future in which every patient's genome will be sequenced upon admittance to the hospital and various sequencing strategies used concurrently to guide the decision-making process.

In order for this future to become a reality a number of challenges must be addressed. First, the sheer size and heterogeneity of the data being produced will require improvements in data management strategy and computing infrastructure (Shilo et al., 2020). The "big data" revolution, that was popular a few years ago, provided the initial impetus for much of the progress in data management and infrastructure that has occurred to date. During this revolution, big data was described as a type of data with

three main characteristics, that is, the data was said to have the “three V’s”. These included: volume (size of the data), variety (many sources and types of data), and velocity (pace at which the data is produced) (Laney et al. 2001). Later, the three V’s were expanded to include a fourth V, veracity, which describe the biases and trustworthiness of the data being produced (Normandeau et al. 2013). Bioinformatics data encompasses all four of these V’s and many of the tools developed to address the challenges of the big data movement can be applied in healthcare (Shilo et al., 2020). Some of these tools include: the introduction of the Hadoop ecosystem, cloud platforms, and the development of multiple related open-source projects (Ahuja et al., 2012, Archenna et al., 2015).

A second challenge that must be addressed before widespread adoption of clinical bioinformatics techniques in the hospital will be the development of robust analysis methods capable of delivering accurate, reliable, and consistent results (Al Kawam et al., 2017). Healthcare is an industry, similar to the airline industry, where making mistakes could have dire consequences and the margin for error is small. Recently, machine learning methods have become more popular among bioinformatics researchers. They have been applied in wide array of different settings such as: genome-wide association studies (GWAS), enhancer prediction, pharmacogenomics, among many others, however; deploying robust and accurate machine learning models comes with a host of challenges and we have not achieved a level of performance that justifies widespread adoption (Al Kawam et al., 2017, Doherty et al., 2018, Kleftogiannis et al., 2016).

Challenge of developing robust machine learning pipelines

Designing robust machine learning pipelines that generalize to unseen data has always been a challenge (Bishop 2006, Harrington 2012). The process of constructing a pipeline in and of itself involves subjectivity. For example, the choice of pre-processing method, machine learning model, performance metric, hyperparameter tuning strategy, all must be considered when constructing a pipeline. (Géron 2019). In the past, to make any meaningful progress often required the guidance of a trained specialist. Within the last decade, automated machine learning (autoML) frameworks have been introduced to simplify this process. These frameworks are designed to search across the various combinations and provide an optimized solution (Feurer, M. et al., 2019). However, many of these frameworks lack interpretability and it is difficult to know why certain pipelines outperform others.

For example, if one or more components within a machine learning pipeline produce uninformative outputs the overall performance can degrade, but determining the root cause of the decreased performance is difficult given the inherent dependency chain of pipeline components (Lourenço, R. et al., 2019). Diagnosing the point of the failure within a learning pipeline is still a challenge and an area of active research (Olson, R.S. et al., 2016, Lourenço, R. et al., 2019, Feurer, M. et al., 2019).

Application of machine learning to the microbiome

Bioinformatics research involving metagenomic data is not as mature as research involving other data sources such as transcriptomics or proteomics. Hence, many of the foundational studies done for other types of NGS data have not been done for metagenomic sequencing data. For example, there are multiple studies exploring the application of feature selection and feature engineering methods to transcriptomic data and their impact on classifier performance, but very few exist for metagenomic data (Saeys, Y., et al., 2007, Hauskrecht, M. et al., 2007). Performing these experiments are necessary and will form the basis for future improvements in the field.

Finally, given all that machine learning has to offer, it is important to note that even the most robust machine learning pipeline cannot substitute for understanding what is actually happening at the biological level. Models are crude exploratory and explanatory tools, but true progress is only possible by studying the system at the lowest level. Combining bioinformatics insights with experimentation is the ideal combination and tools designed with this purpose in mind will be of the utmost importance.

Research objectives

The objective of this dissertation is to first conduct a benchmarking study comparing various feature transformation strategies to 16s metagenomic data controlling for model and disease type. The second objective is to develop a python framework that expands on this theme, providing a tool to select the optimal machine

learning pipeline which includes: pre-processing, feature transformation, and model choice. The final objective is to apply the tool to a broader research study exploring the role of milk derived osteopontin on the gut microbiome.

Dissertation aims

Aim 1: To determine whether applying various feature selection, feature engineering, or feature extraction methods prior to training influences 16s metagenomic classifier performance

The first aim explores whether applying various feature selection, feature engineering, and feature extraction methods to 16s metagenomic data prior to training improves performance. Analysis methodology and benchmarks for metagenomic data aren't as mature as other forms of sequencing data and limited work has been done in this space to date. Chapter III presents evidence that in the case of colorectal cancer metagenomic data, applying these methods could be advantageous in some cases, but not all. In this chapter, I train and evaluate a random forest classifier across multiple colorectal cancer datasets and compare how each method influences performance.

Aim 2: To develop a python package that lets researchers identify which components of a machine learning pipeline contribute most to performance

The second aim introduces an autoML python framework designed for 16s metagenomic data. The framework is capable of selecting the most performant combination of pre-processing technique, feature transformation method, and machine

learning model for metagenomic data. Additionally, the framework will identify which components of the pipeline contribute most to performance. Chapter III will introduce the package then apply it to a colorectal cancer dataset.

Aim 3: To determine the influence of milk derived osteopontin on the gut microbiome

The third aim will present initial efforts to explore the impact of milk derived osteopontin on gut microbiome. Chapter IV will present evidence linking milk-derived osteopontin to alterations in the gut microbiome and the tool introduced in Chapter III is used to identify target bacterial species that could be responsible for the phenotype observed.

CHAPTER II

BACKGROUND

Machine Learning

Machine learning is a type of artificial intelligence in which computers use large amounts of data to learn how to complete a given task instead of being programmed to do so explicitly (Hornby, et al., 2011, Bzdok, et al., 2018). This paradigm differs from the statistical approach in which strict assumptions are made about the data generating process (Serra, et al., 2018). Instead of strong assumptions, machine learning algorithms use a collection of training examples to learn the unknown target function. There are multiple sub-disciplines of machine learning such as: supervised learning, unsupervised learning, reinforcement learning, active learning, etc., but here I describe supervised learning, which is the type of learning used throughout this dissertation, and future references to machine learning will assume this paradigm.

Machine learning algorithms make weaker assumptions compared to the statistical approach and draw potential functions from a hypothesis set H. The learning algorithm, A, then uses the training examples along with the hypothesis set to produce a final hypothesis, which is the trained model that estimates the unknown target function (Figure 2.1). This approach has become popular among bioinformatics researchers because of its' increased flexibility; making it perfect for situations in which explicit modeling is intractable. In other words, the hypothesis set from which potential functions

are drawn is more expressive compared to other approaches which enables these approaches to detect subtle nuances in the data (Bzdok, et al., 2018).

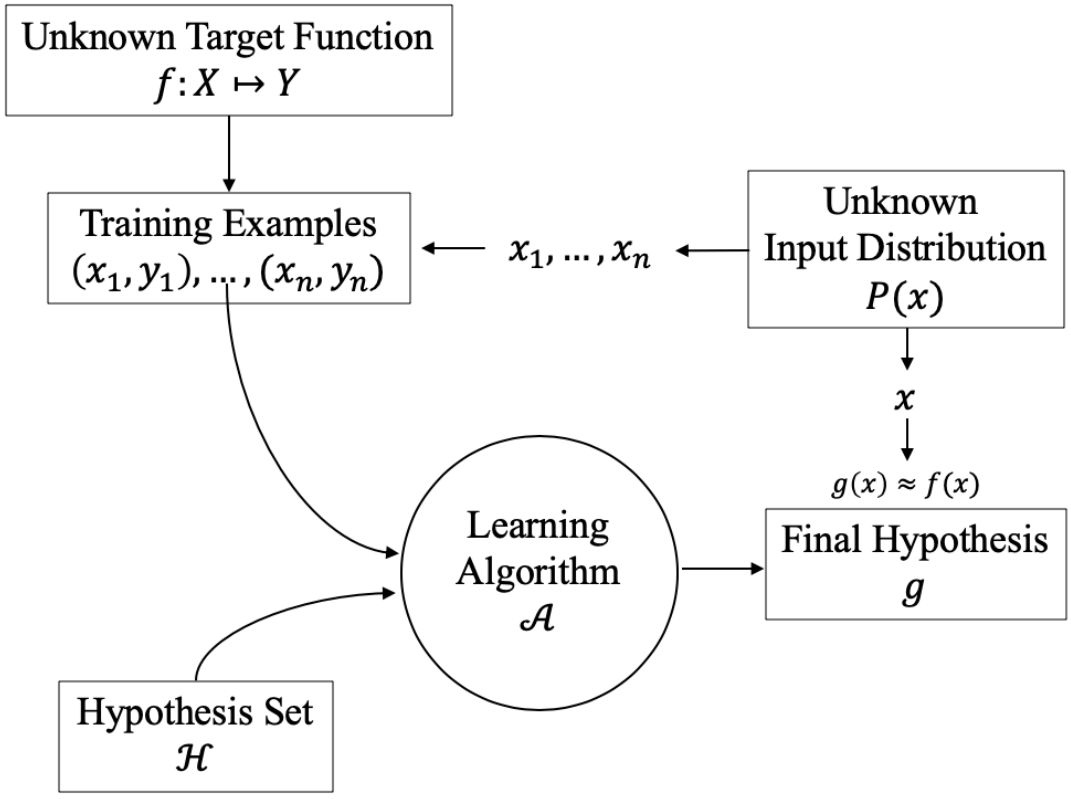

Figure 2.1. Supervised machine learning paradigm

Feature Selection

Machine learning algorithm use training examples to learn an unknown target function. Each training example has a number of attributes, known as features, that describe the data. For example, suppose a researcher would like to use an RNA-Seq

transcription profile to predict a given phenotype, in this case the features would be the genes and their corresponding expression values (Byron et al., 2016, Luo et al., 2017). Feature selection refers to the process of selecting features, or data attributes, that are more relevant for downstream tasks such as classification or regression (Wang, L. et al., 2016). Multiple studies have shown that if a machine learning model is provided with irrelevant features as input, predictive accuracy is reduced, hence studying methods that select informative features is a worthy pursuit (Luo et al., 2017, Wang, L. et al., 2016).

There are three main types of feature selection methods. These include filter, wrapper, and embedded methods. Filter methods simply select features based on some criteria or performance metric. For example, features can be ranked based on their chi-square statistic given the target. Wrapper methods embed the model search space with the feature subset search space in an iterative procedure to eventually arrive at the optimal choice (Figure 2.2). Wrapper methods have the added benefit of including interactions with the model, however, it is easier to overfit using this approach and these methods can be computationally expensive (Saeys, Y. et al. 2007). Finally, embedded methods include the search for features in the classifier construction itself. These methods include interactions, but are less computational expensive when compared to wrapper methods (Lazar, C., et al., 2012, Saeys, Y. et al. 2007).

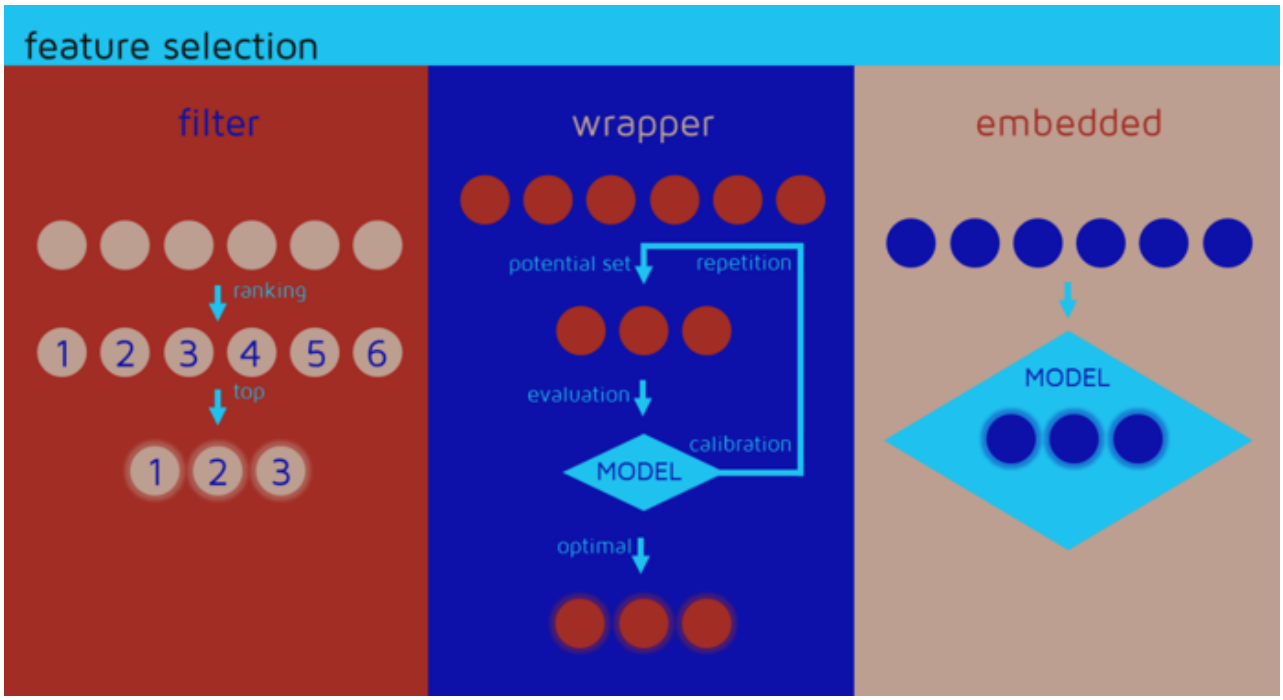

Figure 2.2 Three primary classification for feature selection methods

Feature Extraction and Feature Engineering

Feature extraction is also used to reduce the number of features but the approach taken differs from feature selection. While feature selection selects features from those that already exists, feature extractions creates entirely new features that are derived from the original data (Figure 2.3) (Guyon, I., et al. 2008) Principal component analysis (PCA) is a popular feature extraction technique. The principle components are derived from the original data; however, they differ from the original data and were design to include more information in a compressed format. Although the extracted feature are different from the original input data they can be used as input to the machine learning model (Taguchi, Y. H, et al. 2017). Feature extraction methods seek

to reduce the relevant pieces of information to a smaller set that is used as input to the machine learning algorithm. The term “feature engineering” is typically used interchangeable with feature extraction and that meaning will be assumed through the remainder of this work.

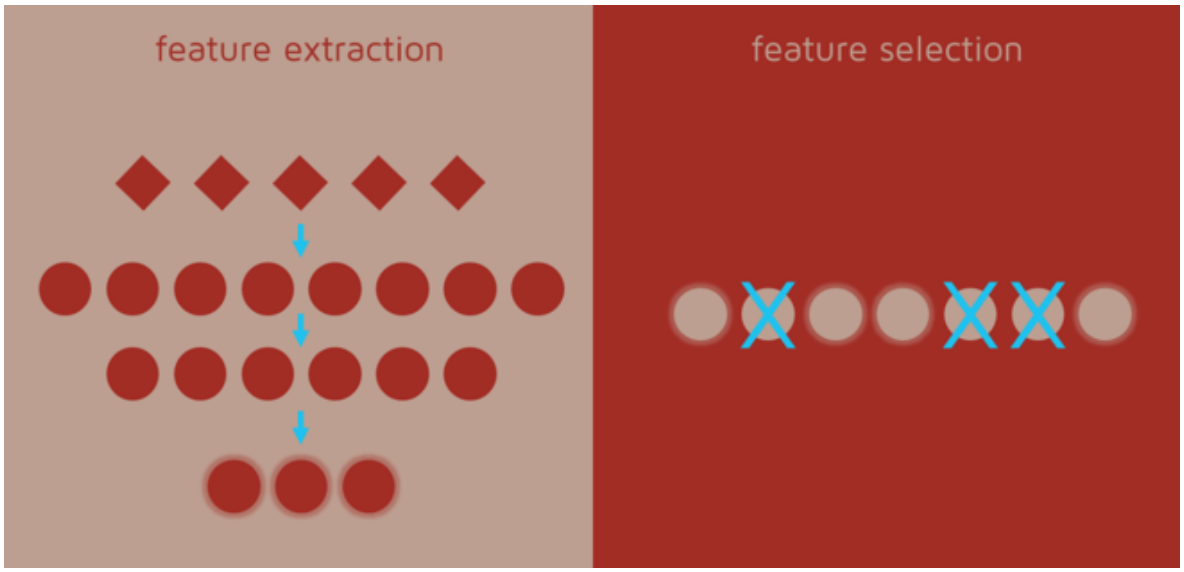

Figure 2.3. Illustration showing difference between feature extraction and feature selection

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) is a technique used to identify groups of genes, or proteins, whose expression is associated with a given phenotype (Subramanian, A et al., 2005). The main idea underlying this technique is to look for overrepresentation of a collection of genes at the top or bottom of a ranked list of genes. More specifically, genes are ranked according to their correlation to a phenotype. Given the ranked list of genes, we can assess whether a collection of genes in a set cluster

towards the top or bottom of the ranked list of genes using a statistical test such as the Kolmogorov-Smirnov (KS) test. It is important to note that the genes within a gene set are usually related in some way and could function together, so detecting cases where a phenotypic class is able to divide out these specific genes could be informative. The gene sets can be visualized using tracks where black bars each represent a gene. In Figure 2.3 we can see a gene set where the black horizontal lines represent genes. Most of the genes in this case cluster toward the top of the ranked gene list where there is high expression in phenotype A compared to phenotype B. Although GSEA is not used explicitly in this dissertation, the underlying principle described forms the basis for how components of a machine learning pipeline are assigned importance.

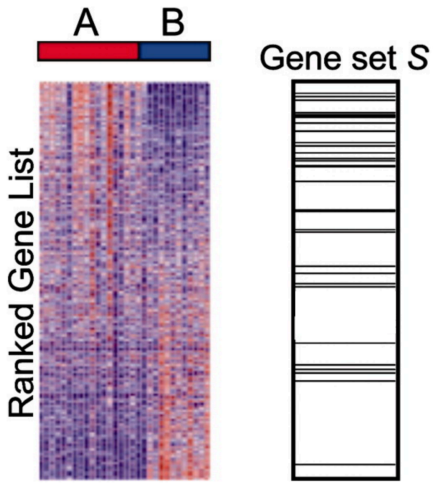

Figure 2.4. GSEA ranked gene list and gene set diagram (Subramanian et al 2005)

Automated Machine Learning

Automated machine learning (autoML) describes an automated procedure for constructing machine learning pipelines. These procedures are able to identify the optimal pipeline end-to-end from raw dataset to complete pipeline meaning it must make decisions regarding data transformation, feature selection or feature engineering, model choice, etc. (Feurer, M., et al. 2019). These procedures make it easier for non-experts to use machine learning on their dataset without extensive background knowledge.

The gut microbiome and 16s metagenomic sequencing

The intestinal microbiome, also known as the gut microbiome, is comprised of the various microorganisms located within the intestinal tract and associated with the host (Arumugam, M., et al., 2011). The gut microbiome plays an important role in many critical functions such as: the proper development of the intestinal tract, and adequate tuning of immunological responses. Moreover, the microbiome is associated with many and varied physiological and disease processes. Therefore, how the microbiome is acquired and maintained are questions that warrant extensive research. 16s metagenomic sequencing is a common sequencing approach to determine the abundance of different bacterial species (Jovel, J et al. 2016). This sequencing strategy relies on the fact that most bacteria share a common 16s gene. This gene has a common and a variable region. The common regions are shared between different bacterial species while the variable regions are different. The variable regions are then

used as a kind of sequencing fingerprint to identify specific bacterial species. The work in this dissertation is based on OTU tables constructed from 16s NGS.

Osteopontin

Osteopontin is a glycosylated phosphoprotein encoded by the Spp-1 gene, originally characterized as part of the rat bone matrix (Frazen, A. et al. 1985, Prince, C.W. et al., 1987). Osteopontin is a pleiotropic molecule involved in a diverse array of physiological and disease processes such as: bone remodeling, atherosclerosis, tumor development and migration, inflammatory bowel diseases, immunomodulation, among others (Di Bartolomeo, M et al., 2016, Giachello, C. M., et al. 1993, Hur, E. M., et al., 2007).

CHAPTER III

LOKKI AN AUTOMATED MACHINE LEARNING FRAMEWORK

Authors

Michael Greer, Zhiao Shi, Bing Zhang

Introduction

The microbiome is defined as the collection of all microbes living inside or on the surface of a host. Recent studies have shown a correlation between microbiome composition and the severity of numerous diseases (Kinros et al., 2011, Qin et al., 2014, Turnbaugh et al., 2006, Zackular et al. 2014). This observation, coupled with the decreasing cost of next-generation sequencing, has driven efforts to develop non-invasive diagnostic models that use microbiome profiles to identify at-risk patients (Bang et al., 2019). However, research involving microbiome-based analysis methods are still very much in their infancy and many of the benchmarking studies done for other types of sequencing data have not been done for metagenomic data.

There are certain characteristics of the human microbiome that make it difficult to analyze. For example, large microbial differences exist between similar individuals, such as twins (Turnbaugh et al., 2009). Microbiome data is also high-dimensional which adds to the complexity. Traditional statistical models do not account for this high level of inter-individual variation and complexity (Segata et al., 2011). This has led many researchers to consider machine learning approaches which are better able to detect differences between samples in spite of high amounts of inter-species variation (Namkung et al.,

2020, Zhou et al., 2019). However, selecting the appropriate data pre-processing method, feature processing method and machine learning algorithm that comprise a complete pipeline is time-intensive and often requires expert knowledge.

Automated machine learning (autoML) frameworks were created to abstract away many of the details required to create machine learning pipelines (Feurer et al., 2015). However, few frameworks have been designed to meet the requirements for microbial classification tasks (Yang et al., 2020). Even fewer frameworks seek to explain the characteristics of top performing pipelines. This exploratory analysis would be useful for a number of reasons such as: determining bottlenecks in performance, identifying cost savings opportunities if a less computational expensive component performs nearly as well as other more performant expensive choice, and gaining insights on how to make future improvements. For example, if non-linear models perform well in general, researchers could focus their attention on these types of algorithms and explore other non-linear models available elsewhere. Hence, there is a need for software that can explain which components of a machine learning pipeline are the most relevant. Here, we first conduct a benchmarking study that explores whether applying various feature selection and feature engineering techniques prior to training influences classifier performance. Next, we expand our study to the development of an entire autoML framework that will simplify the task of selecting pre-processing, feature selection or engineering, and machine learning model for any metagenomic sequencing dataset.

Significance

This work is significant for a number of reasons. First, although comparative analyses showing the effect various feature selection and feature engineering methods have on classification have been done for other types of NGS data, very few have evaluated these methods on 16s metagenomic sequencing data. Metagenomic sequencing data has specific properties that differ from other NGS data sources, so exploring how this data type reacts to various techniques is of interest. Furthermore, in this study we will use 16s metagenomic profiles from colorectal cancer patients which remains a leading cause of cancer-related death. Benchmarking studies are an important first step toward the development of models that can be used in the clinic for early detection of the disease. Another reason this study is significant relates to the development of the autoML framework. This framework is one of the first autoML frameworks designed specifically for metagenomic data. It also introduces a novel visualization technique which enables users to determine which component of a machine learning pipeline are most relevant to performance.

Objectives

There were two objectives of this study. First, to conduct a small benchmarking experiment to determine whether applying a feature selection or feature engineering technique prior to model training improves performance. Second, to develop an autoML framework for 16s metagenomic sequencing data.

Materials and Methods

Data Source

The 16s sequencing data sets were obtained from a recent 16s colorectal cancer meta-analysis (Sze, M. A., et al. 2018).

Construction of model building pipelines

To perform an analysis, users must specify a performance metric then a set of pre-processing, feature transformation, and machine learning algorithms to consider. The package will automatically determine the hyperparameters for all classifiers and then compute performance for all pipelines based on cross validation.

Model building component enrichment analysis

To visualize which components of the model building pipelines contribute most to performance Lokki first sorts each pipeline according to the selected performance metric. Next, it filters the scores based on all combinations of either one (single factor) or two (dual factor) components. Finally, statistical significance is assessed using the Kolmogorov-Smirnov (KS) Test applied to the scores in the current factor or factor combination of interest versus all other scores. The visualization is inspired by gene set enrichment analysis (Subramanian et al., 2005) whereby black bars are drawn to show where in the ranked list of pipelines the factor or factor combination appear.

Pipeline Selection

Users have two options available when selecting a final pipeline for re-use on unseen data. The first option is to select the top performing pipeline based on cross-validation performance. This approach may be optimal in some cases, but may be prone to overfitting. The second option is to walk down the ranked list of pipelines and construct

a model building pipeline by selecting the first of each component type to appear k number of times. For example, suppose k = 3, then if log-transformation is the first pre-processing method that appears three time in the ranked list of pipelines it will be selected, then if PCA is the first feature transformation method to appear three time it will be selected, and so on until all components of the pipeline are complete.

Local Installation

Lokki can be installed on local computers using the python package index (PyPI) by executing *pip install lokki* from the terminal. After installing the package, it can be imported and used without any further setup.

Results

Benchmarking Study

A summary of the datasets, and sample size, for the benchmarking study are provided in Table 3.1. To visualize the amount of diversity within each dataset I measured the alpha diversity using richness (observed OTU) and evenness (shannon score) metrics. There was no significant difference in richness or evenness when comparing normal to cancer patients across all datasets (Figure 3.1)

Table 3.1 Benchmarking study sample sizes

Study	# Normal	# Disease
Ahn	148	62
Baxter	172	120
Flemer	37	43
Hale	473	17
Wang	56	46
Zeller	50	41

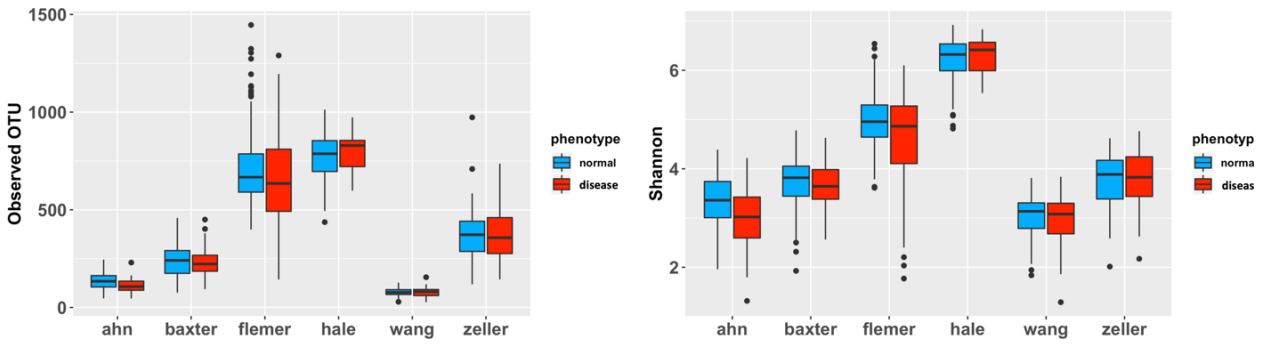

Figure 3.1. Normal vs disease richness (observed OTUs) and evenness (shannon) of microbiome profiles across all datasets

A summary of the feature selection and feature engineering methods for the benchmarking study are provided in Table 3.2 and Table 3.3 respectively. The results of applying these methods prior to training a random forest classifier is provided in Figure 3.2.

Table 3.2 Benchmarking study feature selection methods

Method	Acronym
Chi-Square Statistic	chix
Mutual Information	mi
Minimum Redundancy Maximum Rel.	mrmr

Table 3.3 Benchmarking study feature engineering methods

Method	Acronym
Independent Component Analysis	ica
Kernel PCA	kernel
Principle Component Analysis	pca
Supervised Principle Component Analysis	spca

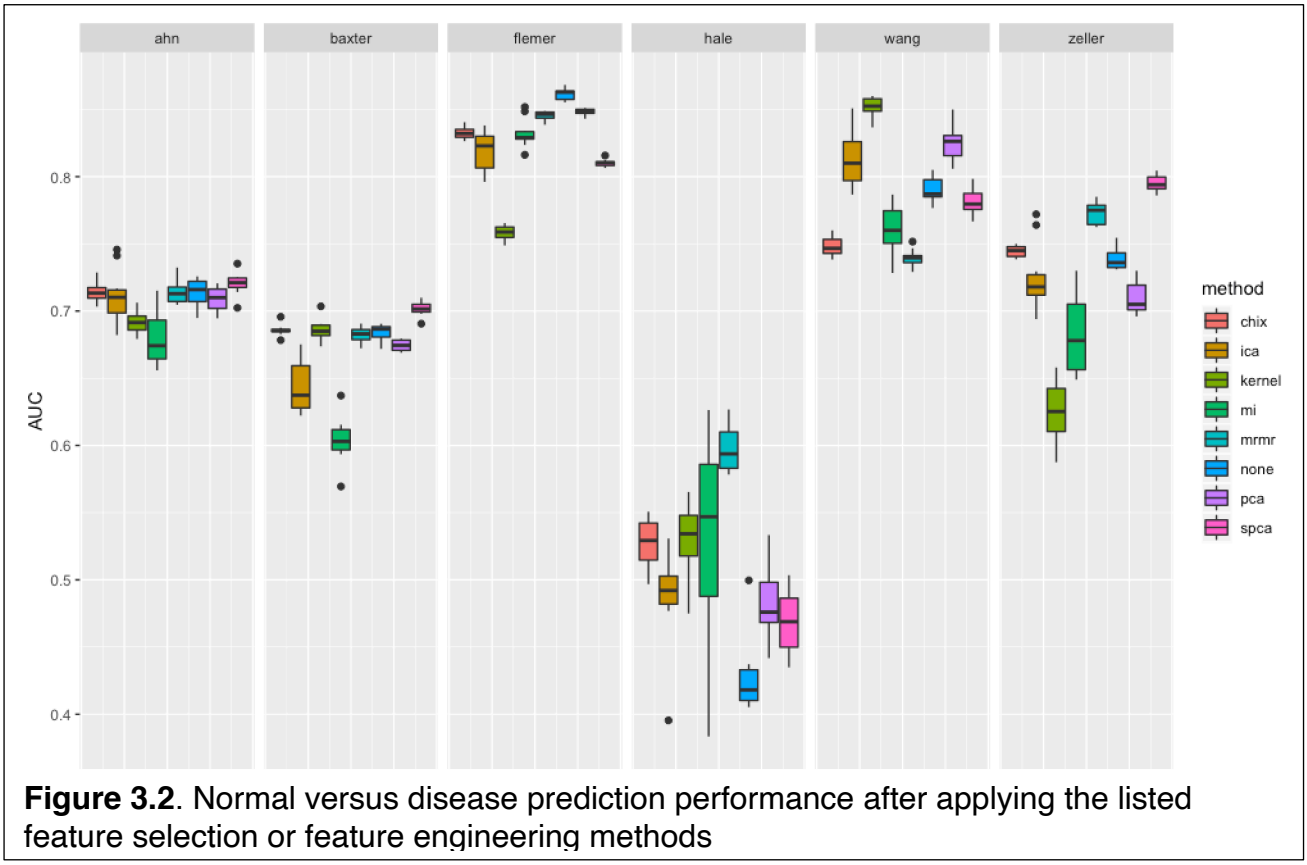

Figure 3.2. Normal versus disease prediction performance after applying the listed feature selection or feature engineering methods

There was not a consistent, or significant, trend observed. In some cases, a feature selection or feature engineering step improved performance and other cases it hurt performance. The results were dataset specific and it is difficult to make recommendations based on the data. A tool is needed to broaden the search to include not only feature selection and feature engineering, but the entire learning pipeline.

Introducing Lokki

Lokki is a python package, designed for metagenomic sequencing data, that allows users to assess which components of a collection of machine learning pipelines contribute most

to performance. A machine learning pipeline consists of three components: a data transformation component, a feature selection or feature engineering component, and a machine learning algorithm (Figure. 3.3).

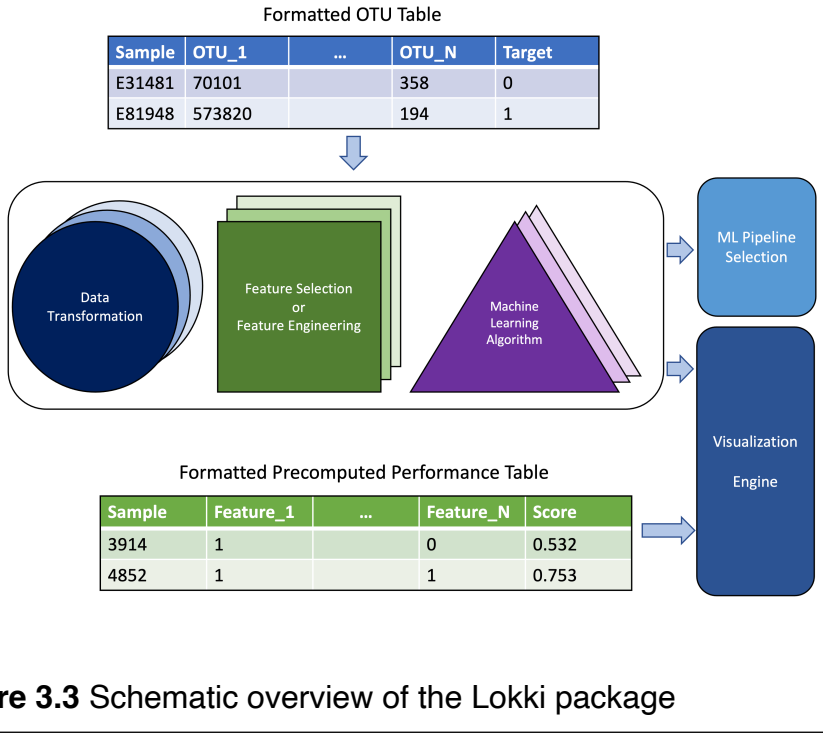

Figure 3.3 Schematic overview of the Lokki package

In addition to typical approaches, such as log-transformation of the data, feature extraction using principal component analysis, and a random forest classification model, Lokki also includes microbiome analysis-specific methods such as taxonomy-aware feature selection (Oudah et al., 2018). A complete list of the methods available for these components can be found in Table 3.4-6.

Table 3.4 Data transformation methods

Name	Description
none	No Data Transformation
log	Log Transformation
zscore	Z Score Transformation

Table 3.5 Feature selection and feature engineering methods

Name	Description
none	No Feature Transformation
hfe	Hierarchical Feature Engineering
chi_square	Chi Square Feature Selection
mutual_information	Mutual Information Feature Selection
pca	Principal Component Analysis
factor_analysis	Factor Analysis
ica	Fast Independent Component Analysis
nma	Non-Negative Matrix Factorization

Table 3.6 Machine learning algorithms

Name	Description
decision_tree	Decision Trees
random_forest	Random Forest
lda	Linear Discriminant Analysis
qda	Quadratic Discriminant Analysis
extra_tree	Extreme Randomized Trees
logistic_regression	Logistic Regression
adaboost	AdaBoost
gradient_boosting	Gradient Boosting
svm	Support Vector Machine
ridge_regression	Ridge Regression

In addition to metagenomic data, users also have the ability to analyze precomputed prediction performance results from a set of machine learning pipelines together with pipeline component information, such as those from the DREAM or other crowdsourcing challenges. In this way, users can identify common characteristics of the winning pipelines (Figure 3.3).

Lokki's visualization engine provides a novel way to compare performance results from all pipelines and to identify key determinants of model performance. A detailed

description of how read the visualization is provided in Figure 3.4. Lokki is a python package designed to function like many other popular data analysis libraries such as pandas and numpy. An overview of how to interact with the package is provided in Figure 3.5-7.

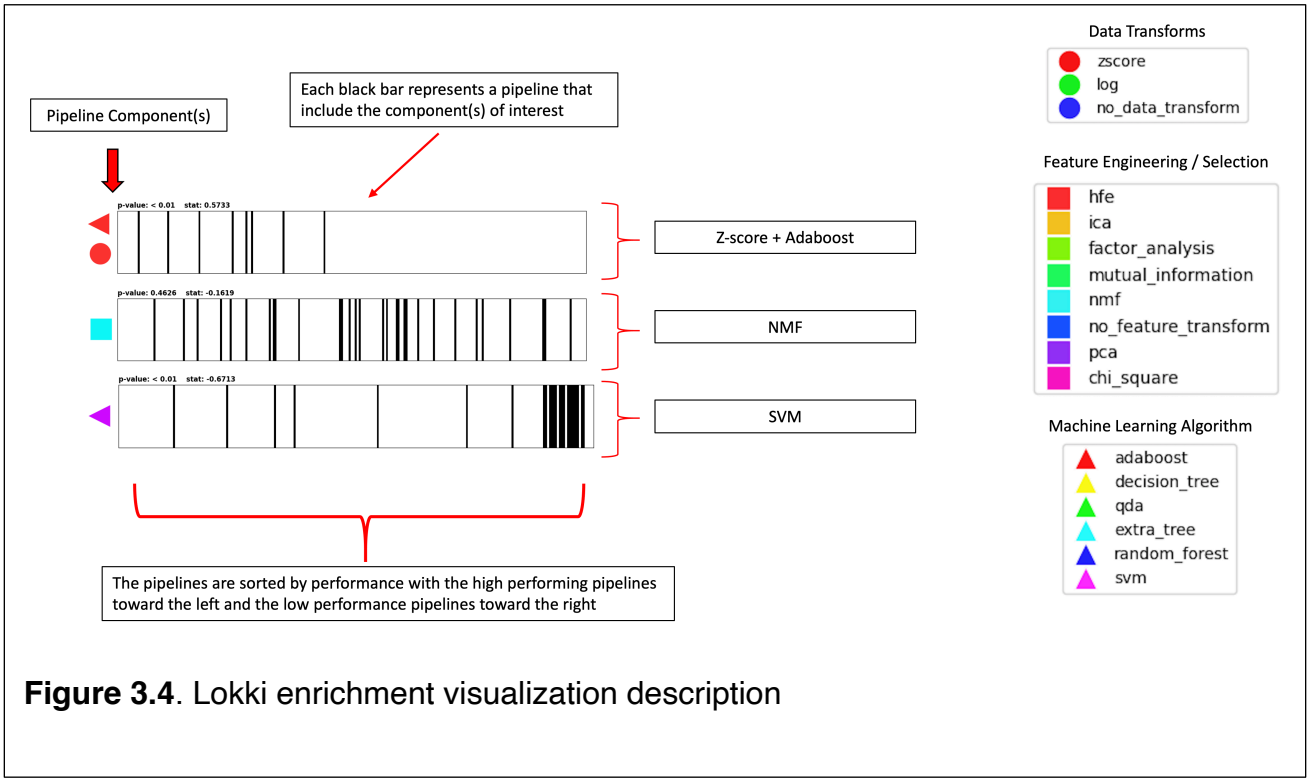

Figure 3.4. Lokki enrichment visualization description

```
import lokki
import pandas as pd

path_to_dataset  = './..'
path_to_taxonomy = './..'

# Load data
data     = pd.read_csv(path_to_dataset)
taxonomy = pd.read_csv(path_to_taxonomy, sep='\t')


# Configure and run
analysis = lokki.configure(dataset = data,
                           target_name = 'target',
                           data_transforms = ['zscore', 'log'],
                           feature_transforms = ['chi_square', 'mutual_information', 'factor_analysis', 'mutual_information'],
                           models = ['decision_tree', 'random_forest', 'ridge_regression'],
                           metric = 'precision',
                           taxonomy = taxonomy)

results = analysis.run()
```

Figure 3.5. Running basic analysis

```
# Output enrichment analysis plots
lokki.plot(analysis_object = results,
           plot_type = 'enrichment',
           filters = ['adaboost', 'logistic_regression'],
           mode = 'dual',
           min_hits = 2,
           num = 20,
           order = 'asc')
```

Figure 3.6. Generating an enrichment analysis plot

```
# Select optimal model
optimal = lokki.select(dataset = data,
                       taxonomy = taxonomy,
                       analysis_object = results)

# Generate prediction on test data
optimal.predict(X)
```

Figure 3.7. Selecting a complete pipeline for later use

To demonstrate the utility of the Lokki software package, we analyzed a previously published 16s dataset from colon cancer patients (Baxter et al., 2016). The OTU table contained 292 patients, 172 normal and 120 disease. Pipeline performances (AUC) ranged from 0.48 to 0.65 (Figure 3.8 top).

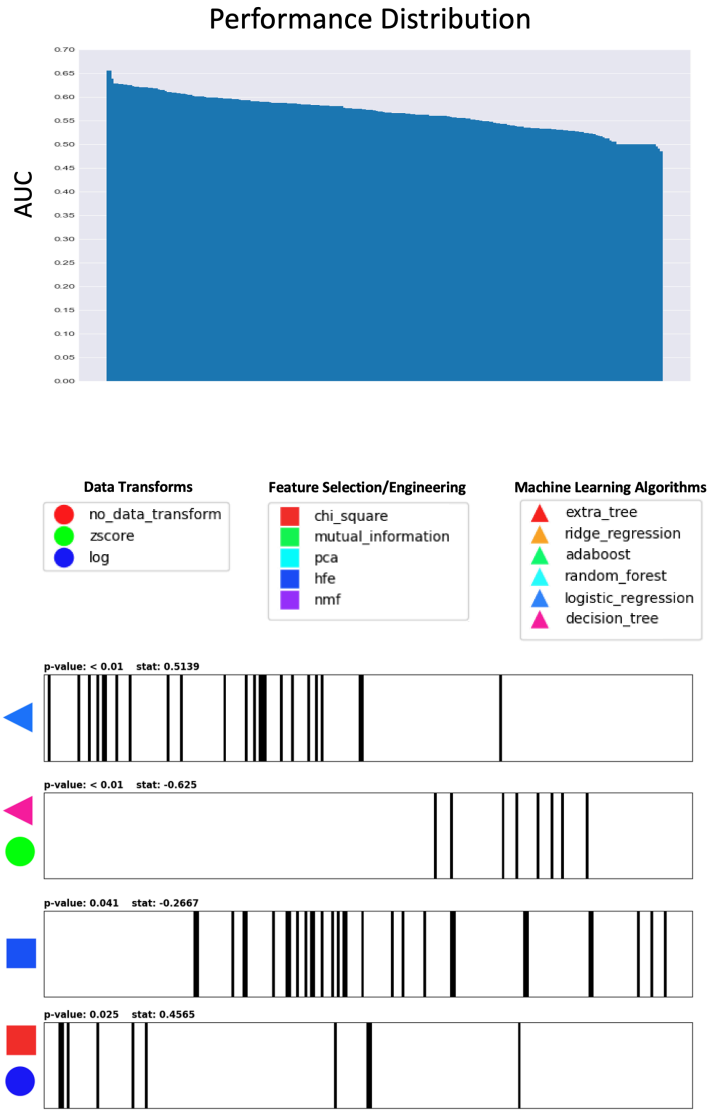

Figure 3.8 Performance distribution and enrichment plot for Baxter dataset

The single factor enrichment analysis results showed no preprocessing resulted in a significant decrease in performance (Figure 3.9). The simple chi-square-based feature selection outperformed other advanced feature selection or feature engineering methods (Figure 3.10).

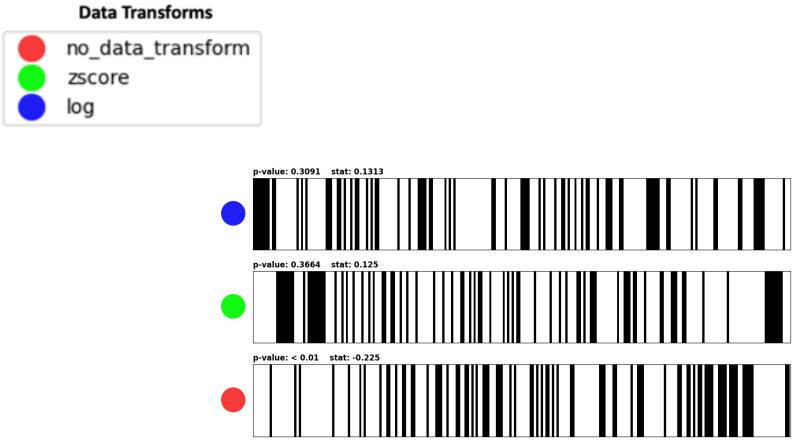

Figure 3.9 Single factor preprocessing results

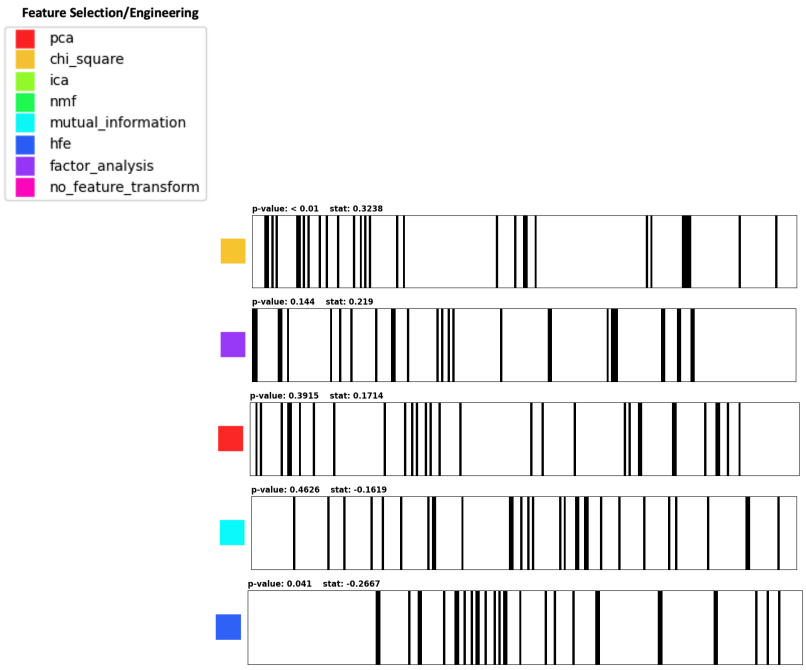

Figure 3.10 Single factor feature selection and feature engineering results

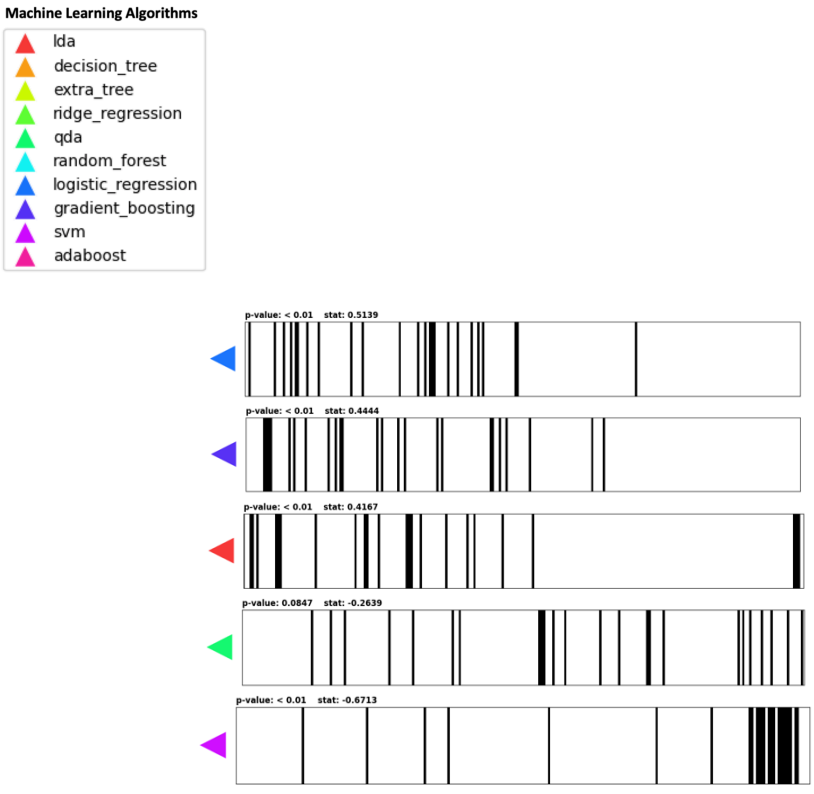

Figure 3.11 Single factor machine learning model results

Interestingly, the microbiome taxonomy-based features did not outperform other approaches in this case despite incorporating taxonomic information (Figure 3.8, 3.10). Among all 10 machine learning algorithms tested, logistic regression showed the best performance, followed by gradient boosting (Figure 3.11). Two factor analysis results revealed more detailed information. For example, chi-square-based feature selection worked well with logistic regression but not SVM (data not shown). This example shows how Lokki can be used to identify key determinants of model performance in a data-driven manner.

Discussion

The benchmarking study did not yield significant or consistent results. In some cases, a feature selection step improved performance while in other cases it decreased performance. Hence, broadening the search to include components of the entire pipeline instead of restricting attention to one component could be advantageous. In response to these results I have developed a software package, Lokki, that helps users discover efficient machine learning pipelines. Lokki provides a novel enrichment analysis method to compare performance results from all pipelines of interest to identify key determinants of model performance. The software was designed with ease-of-use in mind. From a simple installation procedure and data input format to a logical programming interface, this software will be immediately useful to researchers when analyzing new data or precomputed performance results. The enrichment analysis feature presented in this tool

fills an unmet need in predictive modeling of microbiome data, with general applicability to other machine learning tasks.

In order to take full advantage of the Lokki package users may want to consider the following guidelines. First, when performing a de-novo search using an OTU table as input, users should consider filtering out OTU columns that have excessive zeros. For example, users should filter out columns where > 80% of the samples have a zero value before feeding the data into the package. Performing this step will dramatically reduce the time it takes to complete the analysis since metagenomic data is typically zero-inflated and many columns could be eliminated. After running the analysis users may also want to serialize the results object using a package like pickle. This is recommended so the same analysis won't need to be re-run each time the user wants to create a different enrichment plot since the results objects can simply be deserialized. Finally, users should avoid using non-parametric models if they have a small sample since there is an increased probability the resulting models will be overfit. Although there is not a standard method to determine the appropriate sample size given a particular model, there are heuristics users are encouraged to explore and follow.

While applying the tool to our selected dataset the simple feature selection and machine learning methods worked best. There are two primary reasons why the simple feature selection and machine learning model performed better. First, microbiome data is inherently noisy and there is a high level of variance even among the same individual, that is, the data has a low signal to noise ratio. Simple approaches are less likely to fit the noise compared to more complex approaches that have more representational

power. Given the increased representational power, complex approaches can easily fit the noise which would lead to poor performance on out of sample data (i.e. low generalizability) since the patterns detected in the noise are unlikely to repeat.

The second reason has to do with the relationship between the complexity of a model, VC-dimension, and sample size. The VC-dimension is a metric used to characterize the complexity of machine learning models where higher values are associated with more complex models. It has been shown previously that the test error on out of sample data can be bounded by the training error plus another term. More specifically, it has been shown that:

$$\Pr\left(\text{test error} \;\leq\; \text{training error} \;+\; \sqrt{\frac{1}{N}\left[D\left(\log\left(\frac{2N}{D}\right)+1\right)-\log\left(\frac{\phi}{4}\right)\right]}\right)=1-\phi$$

where D is the VC-dimension, N is the sample size and $\phi \in [0,1]$ is constant. Note how the upper bound on the test error shrinks as $N \to \infty$, that is, the term under the square root will tend to 0 x (some large number) = 0 as $N \to \infty$ which means the bound on the test error decreases as the sample size increases. A tight bound would imply similar, or better, test error when compared to the training error (i.e. high generalizability). Also note how as $D \to \infty$ the bound on the test error grows to infinity. That is, as the model choice becomes more complex it becomes harder to get a tight bound on the test error. The data in my study had a small sample size and coupling that with a complex model

would result in a large bound and it is less likely the models will generalize which helps explain the poor performance.

CHAPTER IV

MILK DERIVED OSTEOPONTIN AND THE INTESTINAL MICROBIOME

Authors

Michael Greer, Ali Nazmi, Kristie Hoek, Danyvid Villagomez-Olivares

Introduction

The impact of the microbiome in both health and disease has been revealed in the past two decades. It is now well established that the microbiome has a remarkable influence on obesity, modulation of immune responses, development of autoimmune disorders, among other physiological processes (Valdes, A. M., et al. 2018). Because of its importance, how the microbiome is acquired during birth, how it matures and how it is maintained are important questions that warrant thorough investigation.

One of the most relevant determinants of microbiome colonization after birth is breastfeeding (Rautave, S., et al., 2016). A recent meta-analysis show a significant divergence in the microbiome and its effects in exclusively breastfeeding infants and those that have other nourishment mechanisms (Ho, N. T., et al., 2018). Of its many benefits, there is increasing evidence that breast milk may represent a source of bacteria for colonization of the infant intestines (Urbaniak, C. et al, 2014, Martin, R. et al, 2003), as well as providing complex oligosaccharides to help establishing the nascent microbiome (Rautave, S., et al., 2016). Furthermore, milk also provides important bioactive factors such as immunoglobulins, lysozyme, lactoferrin and defensins that help

the infant battle infections as well as providing a supportive environment for proper microbiome colonization.

An abundant bioactive factor present in milk is osteopontin. Osteopontin concentration in human milk varies depending on whether it is measured in colostrum, early milk (72 h to 7 d post-partum) or mature milk (28 d post-partum), but it ranges from 18 to 322 mg/ml (138 mg/ml in average), which constitutes around 2% of the total protein in human milk (Nagatomo, T., et al., 2004, Schack, L. et al. 2009). The high levels of osteopontin in milk suggest an important role for this protein in the development of neonates and the intestinal microbiome, however, there are limited published works describing the effect of maternal milk-derived osteopontin on the intestinal microbiome.

Significance

The significance of this work resides on the importance of breastfeeding in the development of infants. It is well established that the microbiome is acquired during birth, but there is increasing evidence suggesting that breast milk represents an important source of bacteria for colonization of the infant. In addition to providing nutrients to the neonate, breast milk also contains important bioactive factors, such as osteopontin, that could modulate the nascent microbiome. This study will provide important evidence about the role of milk-derived ostepontin in the development of a proper intestinal microbiome. Completion of this study will also serve as a foundation for future research focusing on how microbiome dysbiosis, caused by insufficient milk-derived osteopontin, impacts the

susceptibility to diseases such as obesity, inflammatory bowel diseases, diabetes, among others.

Objectives

The objectives of this study were two-fold. First, to determine the influence of osteopontin on the development of the gut microbiome. Second, to investigate whether osteopontin influences the growth of bacteria present within the microbiome.

Materials and Methods

Mice

C57BL/6J were originally purchased from The Jackson Laboratory (000664) and have been maintained and acclimated in our colony for several years. Spp-1$^{-/-}$ (004936) mice on the C57BL/6 background were originally purchased from The Jackson Laboratory. Spp-1$^{-/-}$ mice were crossed with C57BL/6J wild-type mice to generate heterozygous offspring, and subsequently bred among themselves to generate knockout mice. Male and female mice were used for all experiments. Mice were maintained in accordance with the Institutional Animal Care and Use Committee at Vanderbilt University.

DNA Extraction and 16s Analysis

Stool was collected from 5 individual WT and and 6 Spp-1$^{-/-}$ mice at three different time points. DNA was extracted using the QIAGEN PowerSoil Kit and processed following manufacturer's instructions. Sequencing was performed on an the Illumina MiSeq

platform (2 x 100 paired-end reads). The python tool QIIME (Caporaso, J. G., et al., 2010) was used for quantifying OTU abundance and computing alpha diversity metrics.

Ensemble feature selection

The autoML package, Lokki, was used to train machine learning pipelines. These pipelines were then filtered for those with AUC > 0.75 and those that included a feature selection step. OTUs were ranked in decreased order of how many separate pipelines the OTU was selected.

Results

To confirm whether or not osteopontin reaches the instestinal tract upon breast feeding we set ♂Spp-1$^{+/-}$ x ♀Spp-1$^{+/-}$ breeders then flushed the small intestines and colon of 21 day and 49 day Spp-1$^{+/+}$ and Spp-1$^{-/-}$ littermates with PBS which was then used on an osteopontin-specific ELISA. Our data indicates that osteopontin is present in the Spp-1$^{-/-}$ mice which confirms the presence of milk-derived osteopontin from the ♀Spp-1$^{+/-}$ dam.

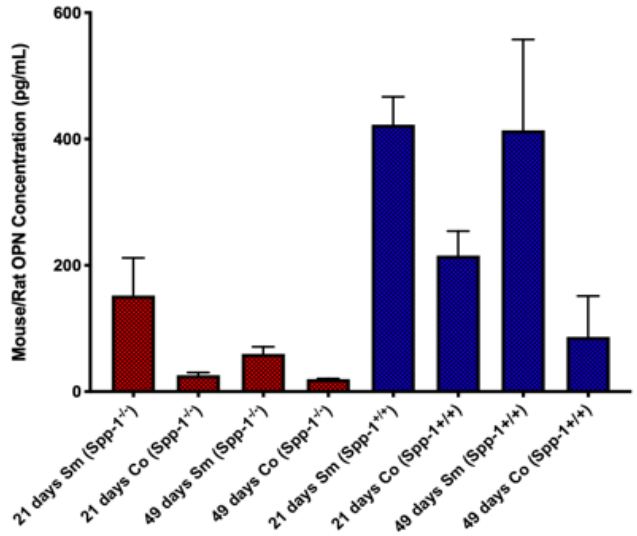

Figure 4.1 Osteopontin concentation in the small intestine and colon of 21d and 49d Spp-$1^{-/-}$ mice and Spp-$1^{+/+}$ mice.

A recent publication, using 8 week old co-housed but not littermates WT and Spp-$1^{-/-}$ mice, showed that osteopontin-deficient mice presented decreased relative levels of the phyla Bacteroidetes, but increased Firmicutes and Proteobacteria (Ito, K., et al., 2017). To corroborate these published results, we set ♂Spp-$1^{+/-}$ x ♀Spp-$1^{+/-}$ breeders and the microbiome from offspring Spp-$1^{+/+}$ and Spp-$1^{-/-}$ littermates was analyzed at 21, 35 and 49 days of age. As shown in Figure 4.2, the observed number of operational taxonomic units (OTU) were similar for Spp-$1^{+/+}$ and Spp-$1^{-/-}$ mice at 21 and 35 days of age. Interestingly, at 49 days of age, the number of observed OTU was significantly higher in Spp-$1^{+/+}$ mice than in littermate Spp-$1^{-/-}$ mice, and correlated with greater microbiome evenness at the same time point (Figure 4.3).

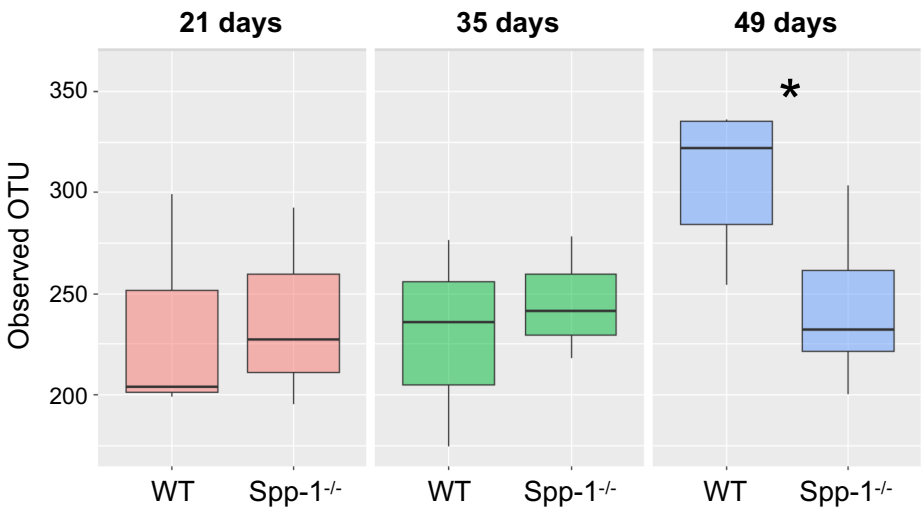

Figure 4.2 Estimated operational taxonomic units present in stool microbiome in the indicated mice at different time points. n=3 to 6 mice. *P<0.05 by Mann-Whitney Test.

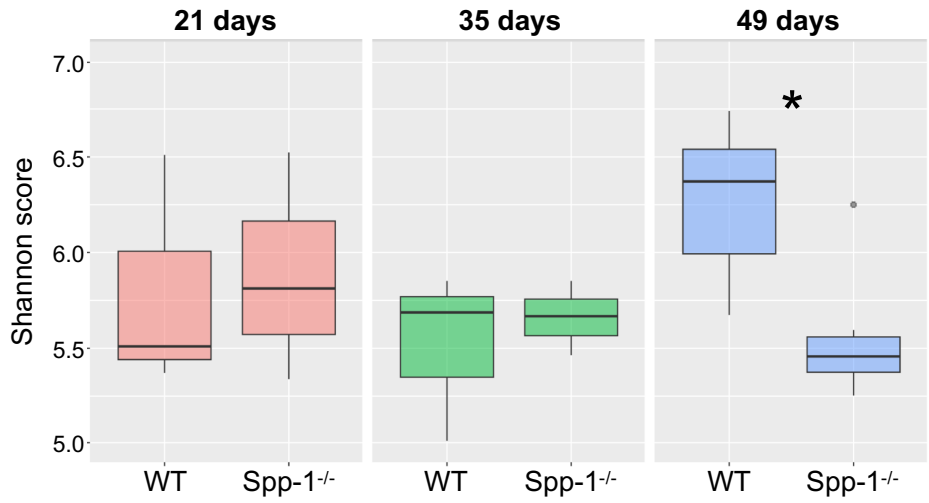

Figure 4.3 Evenness of the microbiome present in stool in the indicated mice at different time points. n=3 to 6 mice. *P<0.05 by Mann-Whitney Test.

The composition of the microbiome at the order level showed primarily an increase in Clostridiales and Desulfovibrionales in Spp-1$^{+/+}$ mice in comparison to littermate Spp-1$^{-/-}$ mice at 7 weeks of age (Figure 4.4). On the other hand, the relative frequency of Bacteroidales and Lactobacillales was higher in 49-day-old Spp-1$^{-/-}$ mice than in Spp-1$^{+/+}$ littermates. We validated the 16S sequencing results by real-time PCR and found similar results. It is important to note that our results differ from those reported by Ito et al., suggesting that the use of littermate animals versus co-housed, non-littermates may be responsible for the observed discrepancy.

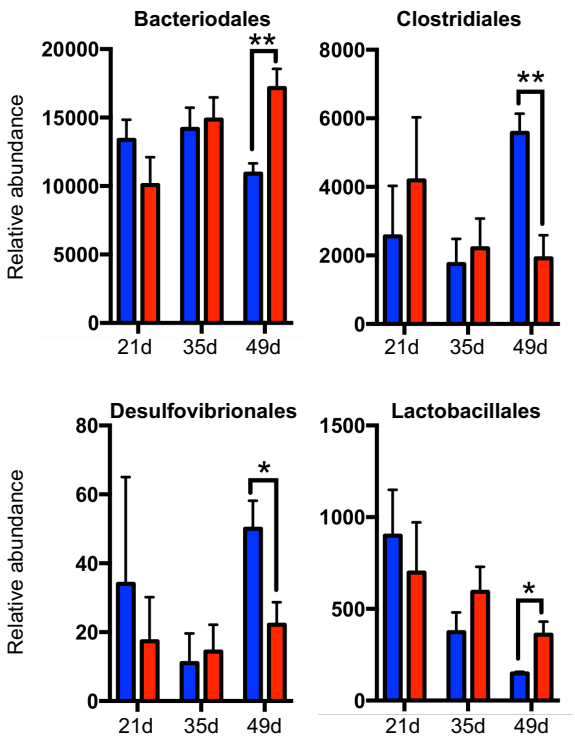

Figure 4.4 Relative stool microbiome order frequencies present in stool. Data derived from 16S sequencing in the indicated mice at different time points. n=3 to 6 mice. Blue bars, wild type mice; red bars, Spp-1$^{-/-}$ mice. *P<0.05; **P<0.01 by Student's t test.

The clostridiales order exhibited the most significant difference between wild type and Spp-1$^{-/-}$ mice. Two bacterial species, clostridioides mangenotti and clostridium innocuum, from the clostridiales order were grown in culture with different concentations of osteopontin to determine whether osteopontin influenced the growth in culture. The results are shown in Figure 4.5. Increasing osteopontin prevented the growth of the bacteria in almost every case.

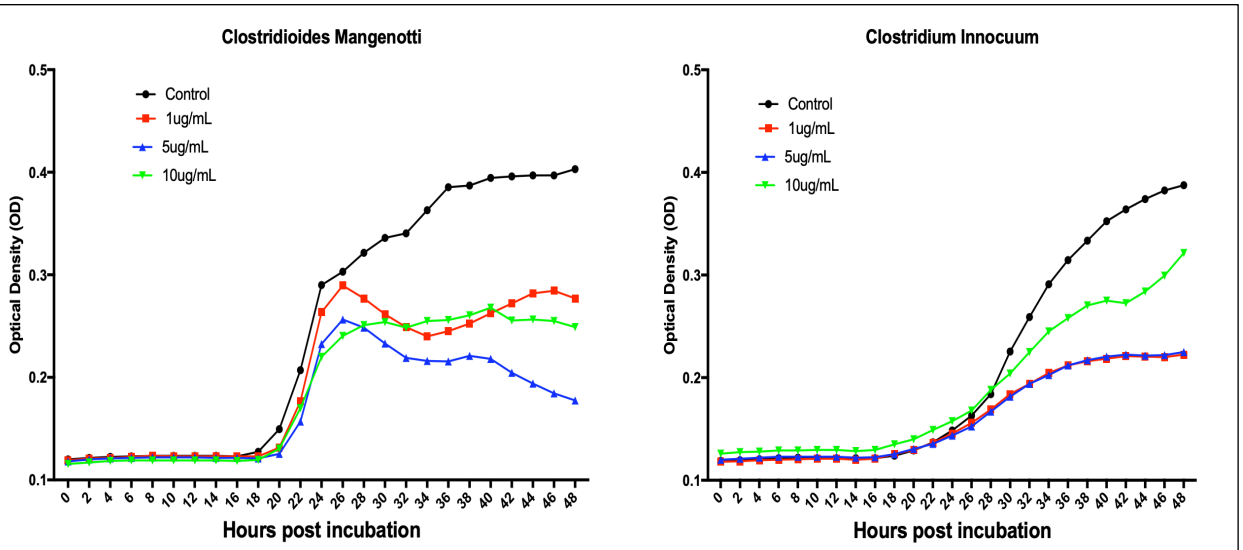

Figure 4.5 Growth rate (by optical density) of clostridioides mangenotti and clostridium innocuum at varying concentrations of osteopontin.

The python package, Lokki, was used to identify other bacteria of interest using ensemble feature selection. The results are summarized in Table 4.1

Table 4.1 Lokki prediction of relevant bacteria

ID	Lokki Prediction	Rank
OTU0011	Deferribacteraceae	1
OTU0024	Lachnospiraceae	2
OUT0009	Prevotellaceae	3
OUT0026	Ruminococcaceae	4
OUT0018	Lactobacillaceae	5

Discussion

Here, we show that mice deficient in osteopontin (encoded by the Spp-1 gene) present gut microbiome dysbiosis. Furthermore, we provide evidence that in Spp-$1^{-/-}$ mice the relative abundance of Bacteriodales and Lactobacillales is increased, whereas the abundance of Clostridiales and Desulfovibrionales is decreased when compared to Spp-$1^{+/+}$ littermate controls. Interestingly, the observed dysbiosis was evident only at around 49 days after birth. Because the littermate mice used for these experiments derived from Spp-$1^{+/-}$ mothers, Spp-$1^{-/-}$ pups were nurtured with osteopontin present in the milk until they were weaned at 21 days of age, at which point the source of osteopontin was eliminated. I also show that osteopontin can influence the growth of bacteria in culture. Lokki was able to predict relevant bacteria for further consideration using ensemble feature selection, however, these predictions still require experimental validation.

I identified Bacteriodales, Clostridiales, Desulfovibrionales and Lactobacillales as relevant bacteria without Lokki while I identified Deferribacteraceae, Lachnospiraceae, Prevotellaceae, Ruminococcaceae, and Lactobacillaceae using tool. However, it's important to note that the bacteria identified without the tool used data that was summarized at the order-level while the data used with the tool were summarized at the family-level. A majority of the bacteria identified using the tool were actually a part of the relevant order-level bacteria. For example, Lachnospiraceae and Ruminococcaceae are both a part of the Clostridiales order which previous work has shown to be correlated with inflammatory bowel disease (IBD) severity. Also, Prevotellaceae is a part of the

Bacteroidales order. A t-test was used to identify the relevant bacteria without the tool. Some benefits of this approach are simplicity, it is computationally inexpensive, and previous work suggests it is a conservative statistical test compared to other methods, however, a negative of this approach is that it makes assumptions about the distribution of OTU abundances which may be incorrect. A non-parametric test could have been used as an alternative. An ensemble feature selection approach was used with the Lokki package. A benefit of this approach is that ensembles are known to decrease variance and increase the stability of predictions which could give more accurate results given the low signal to noise ratio of microbiome data. A negative of this approach is ensembles tend to be more computationally expensive.

CHAPTER V

CONCLUSION

Clinical bioinformatics is a rapidly evolving field that has the potential to transform the healthcare industry. It is likely that future patients will be screened and given personalized solutions for their healthcare needs. This future requires improvements in both computing infrastructure and analysis methodology. The big data revolution has initiated progress in some respects, they've introduced data management software and cloud computing infrastructure capable of managing petabyte-scale data, however improvements in analysis methodology are still needed. Researchers have had success leveraging the power of machine learning algorithms to analyze large quantities of data but developing robust machine learning pipelines is still a challenge.

Creating tools that make it easier to the develop robust machine learning pipelines will be of paramount importance. Automated machine learning frameworks were introduced to ease the pipeline development process and abstract away many of the details required to evaluate, tune, and select models that will generalize out of sample. While many automated machine learning frameworks have been introduced in recent years, most do not assign important to the pipeline components and it is still difficult to know how each components effect performance. This information can help researchers make informed decisions and select for robustness.

This dissertation focused on the development and use of bioinformatics analysis to address two health-related problems using a data-driven and hypothesis-driven

approach. The first problem involved efforts to improve metagenomic sequence-based classifiers performance by conducting a benchmarking study and introducing a novel software package in Chapter III. This type of work is important to advance the development of non-invasive metagenomic-based diagnostics for diseases such as colon cancer. The second problem started with the hypothesis that the protein osteopontin influences the microbiome. The goal of this project was to determine how the absence of osteopontin influences microbiome composition which led to the discovery that milk-derived osteopontin is likely driving some of the observed differences. While completing these studies I learned a number of principles summarized below.

To start, although the benchmarking study results were inconclusive, logical trends were observed in the exploratory analysis of alpha diversity where increased richness and evenness were observed in the normal samples. The results of the benchmarking study also motivated the development of the python-based Lokki software package which is able to construct entire machine learning pipelines and explain which components contribute most to performance. The introduction of this software is a major contribution of my work. The software introduced could have a large impact for a number of reasons. First, it opens up an entirely new dimension of analysis by applying ideas from gene set enrichment analysis to pipeline evaluation. Current autoML approaches completely ignore the wealth of information available by studying how each pipeline performed by component, however, this information can be useful for determining bottlenecks in performance, identifying cost savings opportunities if a less

computational expensive component performs nearly as well as other more performant expensive choice, and gaining insights on how to make future improvements. This tool has the potential to help researcher develop improved models for diagnostics based on a patient's microbiome profile. This is also one of the first autoML packages designed specifically for metagenomic data with microbiome-specific feature selection techniques built into the software.

A central insight from the hypothesis driven study involving osteopontin was that milk-derived osteopontin influences microbiome composition. I arrived at this conclusion after sequencing stool samples collected at various time points and noticing a time dependent difference in the microbiome. Further experiments supported the hypothesis that milk-derived osteopontin is present within the intestine. This result is significant because few studies to date have explored the impact of milk-derived osteopontin on the microbiome. Also, given that infant formula has a much lower concentration of osteopontin than breast milk, and that many mothers use formula as an alternative to feed their children, it is essential to understand the functions of milk-derived osteopontin and its impact in the development of newborns. The results from this study provided the impetus for future work in this direction.

CHAPTER VI

FUTURE DIRECTIONS

A major limitation of the current work was the lack of experimental validation in Chapter IV for the predictions made by Lokki. A number of experiments could have been done for a more detailed understanding of the factors driving the observed phenotypes. For example, the bacterial strains identified from the tool could have be cultured in the presence or absence of recombinant osteopontin. We could then measure the growth rate over time and see whether osteopontin influences their growth. Another experiment could introduce osteopontin to osteopontin-deficient mice through oral gavage in order to see if the wild-type phenotype is rescued. These experiments, and many others, are currently in progress.

REFERENCES

Ahuja, S. P., Mani, S., & Zambrano, J. (2012). A survey of the state of cloud computing in healthcare. Network and Communication Technologies, 1(2), 12.

Al Kawam, A., Sen, A., Datta, A., & Dickey, N. (2017). Understanding the Bioinformatics Challenges of Integrating Genomics into Healthcare. IEEE journal of biomedical and health informatics, 22(5), 1672-1683.

Archenaa, J., & Anita, E. M. (2015). A survey of big data analytics in healthcare and government. Procedia Computer Science, 50, 408-413.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., ... & Bertalan, M. (2011). Enterotypes of the human gut microbiome. nature, 473(7346), 174-180.

Ashley, E. A. (2016). Towards precision medicine. Nature Reviews Genetics, 17(9), 507.

Bang, S., Yoo, D., Kim, S. J., Jhang, S., Cho, S., & Kim, H. (2019). Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. Scientific reports, 9(1), 1-9.

Baxter, N. T., Ruffin, M. T., Rogers, M. A., & Schloss, P. D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome medicine, 8(1), 1-10.

Bishop, C. M. (2006). Pattern recognition and machine learning. springer.

Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., & Craig, D. W. (2016). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nature Reviews Genetics, 17(5), 257-271.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Points of significance: statistics versus machine learning.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... & Huttley, G. A. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature methods, 7(5), 335-336.

Di Bartolomeo, M. et al. Osteopontin, E-cadherin, and beta-catenin expression as prognostic biomarkers in patients with radically resected gastric cancer. Gastric Cancer 19, 412-420, doi:10.1007/s10120-015-0495-y (2016).

Doherty, A., Smith-Byrne, K., Ferreira, T., Holmes, M. V., Holmes, C., Pulit, S. L., & Lindgren, C. M. (2018). GWAS identifies 14 loci for device-measured physical activity and sleep duration. Nature communications, 9(1), 1-8.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2019). Auto-sklearn: efficient and robust automated machine learning. In Automated Machine Learning (pp. 113-134). Springer, Cham.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In Advances in neural information processing systems (pp. 2962-2970).

Franzen, A., & Heinegård, D. (1985). Isolation and characterization of two sialoproteins present only in bone calcified matrix. Biochemical Journal, 232(3), 715-724.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Giachelli, C. M. et al. Osteopontin is elevated during neointima formation in rat arteries and is a novel component of human atherosclerotic plaques. J Clin Invest 92, 1686-1696, doi:10.1172/JCI116755 (1993).

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (Eds.). (2008). Feature extraction: foundations and applications (Vol. 207). Springer.

Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. New England Journal of Medicine, 363(4), 301-304.

Harrington, P. (2012). Machine learning in action. Manning Publications Co..

Hauskrecht, M., Pelikan, R., Valko, M., & Lyons-Weiler, J. (2007). Feature selection and dimensionality reduction in genomics and proteomics. In Fundamentals of data mining in genomics and proteomics (pp. 149-172). Springer, Boston, MA.

Ho, N. T. et al. Meta-analysis of effects of exclusive breastfeeding on infant gut microbiota across populations. Nat Commun 9, 4169, doi:10.1038/s41467-018-06473-x (2018).

Hornby, A. S., & Cowie, A. P. (2011). Oxford advanced learner's dictionary (Vol. 1430). Oxford: Oxford university press.

Hur, E. M. et al. Osteopontin-induced relapse and progression of autoimmune brain disease through enhanced survival of activated T cells. Nat Immunol 8, 74-83, doi:10.1038/ni1415 (2007).

Ito, K. et al. The potential role of Osteopontin in the maintenance of commensal bacteria homeostasis in the intestine. PLoS One 12, e0173629, doi:10.1371/journal.pone.0173629 (2017).

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., ... & Wong, G. K. S. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. Frontiers in microbiology, 7, 459.

Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research, 26(7), 990-999.

Kinross, J. M., Darzi, A. W., & Nicholson, J. K. (2011). Gut microbiome-host interactions in health and disease. Genome medicine, 3(3), 14

Kleftogiannis, D., Kalnis, P., & Bajic, V. B. (2016). Progress and challenges in bioinformatics approaches for enhancer identification. Briefings in bioinformatics, 17(6), 967-979.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70), 1.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., ... & Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(4), 1106-1119.

Lourenço, R., Freire, J., & Shasha, D. (2019). Debugging machine learning pipelines. In Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning (pp. 1-10).

Luo, P., Tian, L. P., Ruan, J., & Wu, F. X. (2017). Disease gene prediction by integrating ppi networks, clinical rna-seq data and omim data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16(1), 222-232.

Martin, R. et al. Human milk is a source of lactic acid bacteria for the infant gut. The Journal of pediatrics 143, 754-758, doi:10.1016/j.jpeds.2003.09.028 (2003).

Nagatomo, T. et al. Microarray analysis of human milk cells: persistent high expression of osteopontin during the lactation period. Clin Exp Immunol 138, 47-53, doi:10.1111/j.1365-2249.2004.02549.x (2004).

Namkung, J. (2020). Machine learning methods for microbiome studies. Journal of Microbiology, 58(3), 206-216.

Normandeau, K. (2013). Beyond volume, variety and velocity is the issue of big data veracity. Inside big data.

Olson, R. S., & Moore, J. H. (2016). TPOT: A tree-based pipeline optimization tool for automating machine learning. In Workshop on automatic machine learning (pp. 66-74).

Oudah, M., & Henschel, A. (2018). Taxonomy-aware feature engineering for microbiome classification. BMC bioinformatics, 19(1), 227.

Prince, C. W., Oosawa, T., Butler, W. T., Tomana, M., Bhown, A. S., Bhown, M., & Schrohenloher, R. E. (1987). Isolation, characterization, and biosynthesis of a phosphorylated glycoprotein from rat bone. Journal of Biological Chemistry, 262(6), 2900-2907.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., ... & Zhou, J. (2014). Alterations of the human gut microbiome in liver cirrhosis. Nature, 513(7516), 59-64.

Rautava, S. Early microbial contact, the breast milk microbiome and child health. J Dev Orig Health Dis 7, 5-14, doi:10.1017/S2040174415001233 (2016).

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.

Sajda, P. (2006). Machine learning for detection and diagnosis of disease. Annu. Rev. Biomed. Eng., 8, 537-565.

Schack, L. et al. Considerable variation in the concentration of osteopontin in human milk, bovine milk, and infant formulas. Journal of dairy science 92, 5378-5385, doi:10.3168/jds.2009-2360 (2009).

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. Genome biology, 12(6), R60.

Serra, A., Galdi, P., & Tagliaferri, R. (2018). Machine learning for bioinformatics and neuroimaging. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(5), e1248.

Shilo, S., Rossman, H., & Segal, E. (2020). Axes of a revolution: challenges and promises of big data in healthcare. Nature Medicine, 26(1), 29-38.

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. Current protocols in molecular biology, 122(1), e59.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545-15550.

Sze, M. A., & Schloss, P. D. (2018). Leveraging existing 16S rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. MBio, 9(3).

Taguchi, Y. H., Iwadate, M., Umeyama, H., & Murakami, Y. (2017). Principal component analysis based unsupervised feature extraction applied to bioinformatics analysis. Comput. Methods with Appl. Bioinforma. Anal, 153, 182.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., ... & Egholm, M. (2009). A core gut microbiome in obese and lean twins. nature, 457(7228), 480-484.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature, 444(7122), 1027.

Urbaniak, C. et al. Microbiota of human breast tissue. Applied and environmental microbiology 80, 3007-3014, doi:10.1128/AEM.00242-14 (2014).

Valdes, A. M., Walter, J., Segal, E. & Spector, T. D. Role of the gut microbiota in nutrition and health. BMJ 361, k2179, doi:10.1136/bmj.k2179 (2018).

Wang, L., Wang, Y., & Chang, Q. (2016). Feature selection methods for big data bioinformatics: A survey from the search perspective. Methods, 111, 21-31.

Yang, F., & Zou, Q. (2020). mAML: an automated machine learning pipeline with a microbiome repository for human disease classification. bioRxiv.

Zackular, J. P., Rogers, M. A., Ruffin, M. T., & Schloss, P. D. (2014). The human gut microbiome as a screening tool for colorectal cancer. Cancer prevention research, 7(11), 1112-1121.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods, 12(10), 931-934.

Zhou, Y. H., & Gallins, P. (2019). A review and tutorial of machine learning methods for microbiome host trait prediction. Frontiers in Genetics, 10, 579.