SIGNAL PROCESSING AND MACHINE LEARNING TECHNIQUES FOR ANALYZING METAGENOMIC DATA

A Dissertation

by

MUSTAFA KAMAL MUSTAFA ALSHAWAQFEH

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Erchin Serpedin
Co-Chair of Committee,	Khalid Qaraqe
Committee Members,	Ulisses Braga-Neto
	Jan Suchodolski
Head of Department,	Miroslav Begovic

May 2017

Major Subject: Electrical Engineering

ABSTRACT

Recent advances in high-throughput sequencing technologies open a new era of genomics studies, called metagenomics. Rapidly, metagenomics has presented itself as the standard approach for characterizing the compositional and functional capacity of microbial communities by direct study of the genetic contents recovered from environmental samples without prior culturing. Although these advancements enable researchers to sequence bacterial populations at a reasonable budget, analyzing these massive metagenomic datasets presents significant challenges. This dissertation presents novel computational tools, based on signal processing and machine learning theories, to enable the investigation of biological systems. Two important research problems are addressed in this dissertation.

The first problem addressed herein concerns the identification of the potential metagenomic biomarkers, which play a critical role in understanding the biological process under study and developing possible therapies. Due to the lack of knowledge of the true biomarkers and a standard assessment methodology, evaluating the quality of the detected markers is challenging. Therefore, we begin by developing an evaluation protocol that mimics the knowledge of the true markers to provide a common ground to compare competing algorithms. Next, a new framework for the biomarker discovery problem based on a low rank-sparse (LRS) decomposition is proposed. The instability of a biomarker detection algorithm renders the identified markers questionable and hinders the translation of these findings into clinical applications. To mitigate this problem, we propose the Regularized Low Rank-Sparse Decomposition (RegLRSD) algorithm. RegLRSD adapts the LRS model to incorporate the fact that irrelevant features are expected to present abundance profiles that do not exhibit a significant variation between samples belonging to different

phenotypes. Integrating this prior knowledge helps to guide the recovery process to more accurate and consistent biological results.

The second research problem addressed in this dissertation concerns the development of a computational framework to enable the translation of the identified markers into clinical applications. Identifying potential biomarkers is the foremost step in the process of understanding the relation between the microbial composition shift due to a certain disease. However, from a practical perspective, the microbial alteration needs to be quantified in a single numerical value, which helps clinicians to measure the disease activity and its response to therapy.

DEDICATION

To my family.

ACKNOWLEDGMENTS

Praise be to God, the Cherisher, and Sustainer of the Worlds. I am very grateful for His infinite blessings and mercy.

I have been supported by many people on the journey to the Ph.D. degree. First, I thank my family for their love, support, and standing by me. I am forever grateful to you!

I extend my special thanks to my advisor Dr. Erchin Serpedin for his constant guidance, encouragement, support, and friendly discussions that we have had during the last few years. I am particularly indebted to Dr. Suchodolski for "opening the door" into this challenging, yet exciting field of research. I am grateful to Dr. Khalid Qaraqe and Dr. Ulisses Braga-Neto for serving on my committee and providing valuable advice.

I would like to thank my colleagues at the Department of Electrical and Computer Engineering particularly, Bilal, Ali, Xu, Xuan, Ahmad who provided a nice and friendly work environment. My special thanks for Bilal who introduced me to the field of metagenomics and for his constant motivation. In addition, I want to thank my friends in the Gastrointestinal Laboratory Anitha and Julia for their generous help and support. I must also thank my friends in College Station for their friendships and the carefree fun moments, which helped me stay sane during difficult times.

I am extremely grateful to my in-laws and extended family for their love, prayers, and confidence in me.

Last but not least, I thank the German Jordanian University for the fellowship for graduate study.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Erchin Serpedin, Professor Khalid Qaraqe, Professor Ulisses Braga-Neto of the Department of Electrical and Computer Engineering and Professor Jan Suchodolski of the Department Small Animal Clinical Sciences.

The two canine with idiopathic inflammatory bowel disease (IBD) datasets (used in Chapter 2 and 3), the dogs with exocrine pancreatic insufficiency (EPI) dataset (used in chapter 3), and the dogs with chronic inflammatory enteropathies (CE) dataset (used in chapter 4) were provided by Professor Jan Suchodolski.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a fellowship from German Jordanian University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Metagenomics

Bacteria are microscopic single-celled organisms that were among the first forms of life that appeared on Earth. Primarily, these microorganisms are named based on their shapes. For examples, bacillus, coccus, and vibrio are Latin names that refer to rod, spiral, and comma-shaped, respectively. Bacterial communities thrive in diverse environments and have been observed in acidic hot springs, soil, radioactive waste [1], deepest portions of ocean [2] and Earth's crust, and inside other organisms. It has been estimated that the total number of bacteria on Earth is approximately 5×10^{30} [3].

The microbiota, a conglomeration of all the bacteria living on/in the human body, play an essential role in defining the health and disease states of the host. In general, these microbial inhabitants outnumber the human's cells and comprise about 150 times more genes than the human genome [4]. Some studies have reported that the microbes outnumber the human's cells by a ratio of 10:1 [5], while others limit this ratio to 1.3:1 [6]. These bacterial groups provide a wide range of metabolic functions that a human body lacks [7].

In general, bacteria inhabitants are organized in communities that comprise a vast number of species with complex relationships including mutualism, competition, parasitism, commensalism, amensalism and neutralism [8]. These interactions can be mediated by natural competition for space and resources or via some symbiotic relationships. For example, substances secreted by one species may be metabolized by another [9, 10]. Additionally, members of bacterial communities can interact indirectly through the immune system [11]. Identifying these interactions is crucial to understand the ecological communities and the underlying regulation activities between microbes. For example, the

depletion of a species may affect other species that depend on it for their survival. As an additional example, the oppositional and symbiotic interactions between species contribute to the development and resistance of pathogens [12].

Even though the bacteria have been recognized as playing a key role in defining the health and disease states, their study has represented a challenge in the past due to several reasons. First, the bacteria were mainly studied through cultivation. Many bacterial groups were neither known earlier nor cultivated in a large number in a laboratory setting. For example, the results in [13, 14] estimate that more than 90% of microbes are unknown and uncultivable, while other studies reveal that this ratio may be up to 99% [15, 16]. Second, *in vitro* measurements do not match real *in vivo* values because the laboratory conditions do not match the environment of the host [8]. Recent advancements in sequencing technologies have overcome these limitations and provided researchers with the taxonomic composition and functional capacity of microbial colonies [7, 17].

At present, the sequencing technologies enable the exploration of microbial communities, opening a new era of genomics studies, called metagenomics [18]. Metagenomics is defined as the study of the genetic contents recovered directly from environmental samples without prior culturing. In analogy to the term genome which describes the entire genetic material of a single organism, a metagenome represents the collective genetic material of all the organisms present in an environmental sample. The term metagenomic was first introduced in 1998 by Handelsman et al. [19] to describe the entire genomes of soil microflora.

1.1.1 Metagenomic Techniques

Currently, there are two main kinds of metagenomic studies for analyzing microbial communities: (i) targeted gene metagenomics, and (ii) whole genome metagenomics. In whole genome metagenomics studies, the entire DNA material is subject to shotgun se-

quencing [20]. On the other hand, in targeted gene metagenomics, only certain genes that can serve as phylogenetic markers are sequenced. Specifically, small subunits of the ribosomal RNA (rRNA) (e,g., 16S rRNA for archaea and bacteria or 18S rRNA gene sequences for eukaryotes [21]) are used as phylogenetic markers to determine the taxonomic composition of a sample. rRNA studies provide a cheap and fast way for community profiling. However, metagenomic studies go beyond the compositional information to provide information about the functional capacity of the microbial communities in the sample. Even though the rRNA-based studies consider only specific genes, they are also considered to be "metagenomic" because they analyze an environmental sample with heterogeneous DNA.

1.1.1.1 Targeted Gene Metagenomics

Genetic taxonomic profiling relies on the fact that each organism presents essential genes, which are critical for its survival. The ribosomal RNA genes have been considered as the gold standard for molecular taxonomic classification for decades [21, 22]. Among the variety of rRNA genes, the 16S rRNA is the widely used gene for characterizing the bacterial community composition. This is because of two main aspects of the 16S rRNA gene. First, it is omnipresent among bacteria. Second, the 16S rRNA gene contains nine hypervariable regions ($V1 - V9$). These regions demonstrate significant diversity among taxa, which in turns enables distinguishing species efficiently [23].

In 16S rRNA studies, the targeted genetic region from the extracted DNA from the sample is amplified and afterward sequenced. This amplification process is achieved by means of polymerase chain reaction (PCR). In its essence, PCR is based on the ability of special enzymes called "DNA polymerase" to synthesize a DNA strand that matches the template one. Since the DNA polymerase can add nucleotides only to an existing $3'$-OH group, short stretches of single-stranded DNA that are complementary to the target sequence, called primers, are needed. At the beginning of the reaction, the two strands of

DNA are separated by applying high temperature to the original DNA molecules. Then, primers anneal to the target genes, which enables the DNA polymerase to build the complementary DNA strand, creating a double helix. Repeating this cycle generates a large number of copies of the template DNA. As seen in the PCR process, the starting and the end of the target region have to be known in advance in order to generate the corresponding primers. In practice, those primers can be designed for either single species or whole taxonomical groups.

The targeted gene metagenomics presents two main advantages for taxonomic classifications over the whole genome metagenomics. First, the amplification process of the targeted-gene reduces the number of reads required for accurate taxonomic profiling. Second, since only the essential genes are sequenced, the sequencing cost is much cheaper than that of the metagenomic studies.

On the other hand, 16S rRNA metagenomics presents potential biases. The inconsistency of primers covering different branches of the taxonomic tree is the main source for bias in PCR experiments. In particular, the amplification phase using PCR requires that the targeted genetic region be already known in order to design the appropriate primers. The lack of knowledge of the required primers results in nonuniform coverage of taxa. Chimeric reads is another source of bias. In chimeric reads, the DNA sequences is composed of DNA from two or more organisms, which prevents the correct classification of sequences. An additional major source of bias is the repetition of the essential genes in the genome. For example, the rRNA genes can vary from 1 to 15 copies [24]. This bias can be corrected if the copy number variation is known.

1.1.1.2 Whole Genome Metagenomics

In a whole metagenomic study, the extracted DNAs from the environmental sample are subject to shotgun sequencing, in which the entire genomes present in the sample are

fragmented into small pieces, and then are sequenced. Those small reads belong to coding as well as noncoding regions of all microbial genomes present in the community. Hence, information about the taxonomic composition and the functional capacity of the microbial sample can be gained from those reads.

While this technique counteracts the bias associated with the PCR process, it presents its own sources of biases such as the bias resulted from the assembly process. For example, assembling reads from different species into one sequence renders the correct taxonomic classification of this read impossible. Increasing the length of reads helps in reducing possible assembly errors. Another possible source of bias is the variation in the length of different genomes. Microbial genomes vary from 0.2 Mbp [25] to 10 Mbp [26]. Longer genomes generate more reads which may skew the taxonomic profiles to microbes with longer genomes. Plasmids, small circular double-stranded DNA molecules that are often transferred between bacteria, may introduce errors if the reads generated from plasmids are classified into species different from those to which the plasmids belong to. Similar to the targeted-gene studies, copy number variations lead to biased results with preference to genes with a higher number of copies.

1.1.2 Large-Scale Microbiome Projects

Metagenomics presents itself as a promising approach for investigating microbial communities. Shortly after the invention of metagenomic techniques, various environmental metagenomic studies started. For example, the presence of viruses in seawater was studied in [27]. In an attempt to discover the marine microbial communities, Craig Venter announced the Global Ocean Sampling project in 2004. In this project, the Venter's personal yacht was equipped with DNA sequencing lab to sequence samples from seawater around the globe [28, 29]. As an additional example is studying the bacterial populations inhabiting an acid mine drainage system [30]. In relation to human microbiome, the oral

microbiota [31], the intestinal bacterial composition [32], the distal gut flora [7], and the predominance of the gut microbiome [33] represent few examples of metagenomic studies.

Due to the rapid advancements in metagenomics and their wide applications in other fields including (ex., medicine, food safety, and wastewater treatment), various large scale well-organized projects that incorporate several disciplines were launched. In general, these projects are of global nature in the sense that it connects researchers from around the word. A leading project in this field is the Human Microbiome Project (HMP) [34], which tries to characterize all possible microbial communities of the human body and find a possible relation between alteration of the human microbiome and certain diseases. Metagenomics of the Human Intestinal Tract (Meta-HIT) [4] is another project that aims to find possible associations between human microbiota and his/her health status. Specifically, Meta-HIT focuses on two disorders: (i) inflammatory bowel diseases, and (ii) obesity. The Earth Microbiome Project (EMP) is another promising project, which targets analyzing the microbial compositional and functional diversity across the whole globe.

1.2 Metagenomic Biomarker Discovery

Recently, several metagenomic studies have revealed that the distortion of the normbiosis state of bacterial communities is a key player in the progression of many diseases such as obesity [35, 36, 37], diabetes [38], inflammatory bowel disease (IBD) [39], and cancer [40, 41]. These findings suggest using microbes as possible biomarkers for several host's health and disease states.

Biomarker detection presents itself as a major means of translating metagenomic data into clinical practice [42]. Identifying potential biomarkers is essential in understanding disease evolution and designing antibiotic and/or probiotic therapies. Microbial biomarker discovery aims to identify the specific operational taxonomic units (OTUs), whose relative abundances differ between different phenotypes. Mathematically, identifying biomarkers

is cast as the problem of finding the most informative variables (or features) that discriminate two or more groups of samples (i.e., healthy versus diseased, or different disease stages).

1.2.1 Related Work

The methods proposed in the literature to address the biomarker discovery problem from metagenomic data can be classified into two categories: **statistical methods** and **machine learning methods**.

In general, the statistical methods tackle the problem by using a statistical hypothesis test to calculate the statistical significance (i.e., p-value) of each feature. Then, the features with p-values less than a predefined threshold are selected as potential biomarkers. A major issue associated with the statistical-based methods is the multiple comparisons problem, which is commonly solved by replacing the p-values with the corresponding false discovery rates (FDRs). Metastats [43] and LEFSe [42] are the current standard methods that belong to this category. Specifically, Metastats utilizes the permutation t-test and the exact Fisher's test for non-sparse and sparse features, respectively [43]. On the other hand, to improve the robustness of biomarker discovery, LEFSe couples the statistical analysis with the effect size estimation [42]. In particular, LEFSe employs the non-parametric Kruskal-Wallis and Wilcoxon-Mann-Whitney tests for class and subclass comparisons, respectively.

From the machine learning perspective, the biomarker detection task is formulated as a feature selection problem. The filtering methods are the most widely adopted approaches for biomarker detection. In filtering methods, each OTU is assigned a score based on the relevance between its abundance levels across the samples and the class labels of the samples. The OTUs with highest scores are selected as potential biomarkers. This scoring process is carried out individually and independently of the other OTUs. Therefore, fil-

tering methods are computationally fast and easily interpretable. However, the individual ranking ignores the inter-dependencies among different variables.

Contrary to the individual ranking, the feature transformation-based methods attempt to generate more informative features where each new feature is a function of all the original features. Considering all the initial features in the construction of new features accounts for the interactions between OTUs. However, this transformation process results in losing the interpretation. Transformation approaches are divided broadly into two categories based on whether the labels of the samples are considered in the transformation process. These categories are the supervised and unsupervised methods. Partial least squares (PLS) and linear discriminant analysis (LDA) are the two widely used supervised methods. On the other hand, the principal component analysis (PCA) presents itself as the most prominent unsupervised method. These methods have been extensively used for the analysis of biological data.

The current machine learning-based state-of-the-art method that is designed specifically for metagenomic biomarker detection is MetaBoot [44]. Basically, MetaBoot combines minimal redundancy maximal relevance (mRMR) feature selection method [45] with bootstrapping in order to obtain a non-redundant subset of potential markers.

1.2.2 Challenges Associated with Metagenomic Biomarker Discovery

Identifying the most discriminating features in metagenomic datasets is a challenging task for several reasons. First, the number of features representing potential biomarkers is large, a challenge that is commonly referred to as 'the curse of dimensionality'. The challenges associated with analyzing the high dimensional metagenomic data set is compounded by the small number of available samples. This high-dimension small-sample challenge raises serious analytical challenges [46, 47]. Second, many microbial populations exhibit a high inter-subject variability. For example, [36] shows that the gut bacterial

ecosystems of twins differ significantly. This inter-subject variability adds more confounding factors that complicate the analysis and interpretation of the results. Third, the microbial communities exhibit a high dynamics due to the complex relationships between its members [8, 9, 10] and the direct interaction with the host [11]. Fourth, metagenomic data are subject to their own artifacts including sequencing errors and chimeric reads [48, 49].

These challenges lead to a serious inconsistency problem that prevents many biomarker detection algorithms from identifying the correct biomarkers involved in the biological process under study. For example, the authors of [50] reported that out of the 70 genes that were suggested as potential biomarkers for breast cancer by the two gene expression studies [51, 52], only three genes were found to be common. As an additional striking example, the authors in [53] reported that out of the 207 detected markers in relation to breast cancer from 15 mass spectrometry studies, only 10 of them were in-common in more than 2 studies. Therefore, developing a robust biomarker detection algorithm that ensures the reproducibility of the results drawn from biological data is crucial to derive solid biological conclusions and translate these findings into clinical applications.

In addition to the technical challenges associated with identifying potential markers, the assessment of the detected markers presents its own concerns. In general, performance evaluation is the foremost step once a new technique is developed. Usually, competing techniques are evaluated using benchmark datasets to be fairly compared based on a common ground. Unfortunately, the field of metagenomic biomarker discovery lacks such benchmark datasets to objectively assess the performance of marker selection algorithms. Therefore, the evaluation criteria and comparisons need to be suitably designed in order to mimic the knowledge of true markers.

1.3 Dysbiosis Index

Indeed, identifying potential biomarkers is crucial to characterize the bacterial groups that may explain the systematic imbalance in the bacterial populations between samples belonging to different phenotypes. However, to translate these findings into clinical applications, it is required to integrate the detected markers into an easy-to-use computational framework to measure the disease activity and to measure the response to therapy. In particular, the microbial changes need to be quantified in a single numerical value called the Dysbiosis Index (DI).

In this dissertation, the focus was on developing such a numeric index specifically for IBD. IBD is a complex immunological disorder of the gastrointestinal tract triggered by an abnormal response of the immune system [54, 55, 56]. In particular, the host's immune system mistakenly treats the bacteria and food in the intestinal system as invading substances. This result in an accumulation of white blood cells at the lining of the gut producing inflammation. According to the Crohn's and Colitis Foundation of America (CCFA) annual report published in 2014, there are 1.6 million IBD patients in the United States, with annual increasing rates approximately equal 70000 patients. Moreover, the annual financial burden of IBD in the US was estimated to be $\$14.6 - \$31.6$ billion.

1.3.1 Related Work

Several clinical indices were proposed to measure the disease activity in patients with IBD. These methods differ in the mathematical formula and the predictor variables that are used to generate the index value. Some of these indices include primarily clinical parameters such as stool consistency and frequency, weight loss, the degree of abdominal pain, vomiting. The Crohn disease activity index (CDAI) [57] represents an example of this kind of indices. In addition to the clinical parameters, other indices incorporate objective laboratory indicators of inflammatory activity for computing the scoring index.

For example, the pediatric CD activity index (PCDAI) [58] includes laboratory variables such as hematocrit, erythrocyte sedimentation rate, and serum albumin. As additional examples, the two widely accepted indices for IBD or CE in canine are the canine IBD activity index (CIBDAI) [59] and the canine CE clinical activity index (CCECAI) [60].

Due to the increasing number of metagenomic studies that have associated the imbalance in the gut microbiota with the chronic enteropathies (CE), developing a new clinical activity index for CE that incorporates the abundance levels of certain microbes can be applied to predict the outcome of CE. The authors in [61] proposed the Microbial Dysbiosis index (MD-index), which is defined as the logarithm of the ratio between the total abundance in organisms increased in CD and the total abundance of organisms decreased in CD. Mathematically,

$$\text{MD-index} = \log \frac{\sum_{i \in I_{inc}} d_i}{\sum_{i \in I_{dec}} d_i}, \tag{1.1}$$

where d_i represents the abundance level of the i^{th} bacterial group. Variable I_{dec} denotes the set of microbes that exhibit decreased abundance levels in samples with IBD, whereas I_{inc} represents the set of enriched microbes in IBD subjects. Based on the results in [61], I_{inc} includes: Enterobacteriaceae, Pasteurellaceae, Fusobacteriaceae, Neisseriaceae, Veillonellaceae, and Gemellaceae. On the other hand, I_{dec} includes: Bacteroidales, Clostridiales, Erysipelotrichales, and Bibidbacteriaceae. Another diagnostic test using fecal samples applicable for both IBD and irritable bowel syndrome (IBS) was proposed in [62] and it is referred to as the 'GA-map test'. GA-map test requires DNA probes for 54 bacteria at different taxonomic levels.

1.4 Main Contributions of this Research

The original mathematical and computational contributions of this dissertation are next described.

1.4.1 Chapter 2: Matrix Decomposition Framework for Identifying Potential Metagenomic Biomarkers

It can be seen that there is a lack of standardized methodology for performance evaluation and in many cases, performance evaluation is not even formally addressed in the state-of-the-art metagenomic biomarker detection algorithms. This little attention renders the current evaluation metrics insufficient to assess key performance aspects expected out of a biomarker detection algorithm. For example, a biomarker selection algorithm is expected to provide consistent results irrespective of a small variation in the input dataset. Otherwise, researchers will not be confident about the detected markers. This reproducibility performance is not addressed in the existing metagenomic biomarker detection algorithms. This chapter presents an evaluation protocol which provides a fair and an accurate assessment of the efficiency of a biomarker detection algorithm in terms of both (i) the reproducibility of the detected biomarkers and (ii) the classification performance. In addition to the evaluation protocol, Chapter 2 proposes a low rank-sparse (LRS) matrix decomposition framework for the biomarker detection problem. This formulation allows to re-cast the biomarker detection problem as the conventional robust principal component analysis (RPCA) problem, which can be efficiently solved. The contributions of this chapter is based on the work represented by the following papers:

- **Alshawaqfeh M**, Bashaireh A, Serpedin E, Suchodolski J. Consistent metagenomic biomarker detection via robust PCA. Biology Direct. 2017 Jan 31;12(1):4.

- **Alshawaqfeh M**, Al Kawam A, Serpedin E. Sparse-Low Rank Matrix Decomposition Framework for Identifying Potential Biomarkers for Inflammatory Bowel Disease. Submitted to the European Signal Processing Conference (EUSIPCO 2017).

1.4.2 Chapter 3: Incorporating Prior Knowledge for Improving Metagenomic Biomarker Discovery

In order to enhance the reproducibility performance of the LRS matrix decomposition-based biomarker detection algorithm, a novel Regularized Low Rank-Sparse Decomposition (RegLRSD) algorithm is proposed in Chapter 2. The essence of RegLRSD is to incorporate the prior knowledge that the irrelevant microbes are expected to be uniformly abundant among samples belonging to different phenotypes. The convex formulation of the LRS model of the biomarker detection algorithm provides a natural way to integrate prior knowledge. In particular, smoothness constraints can be imposed over the recovered low-rank matrix that represents the profiles of the non-informative bacteria.

Incorporating the prior knowledge in the biomarker discovery process is a major means to mitigate the instability of marker selection algorithms [63]. The reasoning behind this is that the prior knowledge guides the algorithm to yield more accurate results. To solve this matrix decomposition problem, an efficient solution based on the alternating direction method of multipliers (ADMM) is developed. Comprehensive comparisons with the existing state-of-the-art algorithms are conducted over three realistic datasets. It is shown that RegLRSD outperforms the competing algorithms both in terms of classification accuracy and reproducibility performance. The contributions of this chapter is based on the following work submitted for publication:

- **Alshawaqfeh M**, Bashaireh A, Serpedin E, Suchodolski J. Reliable Biomarker Discovery from Metagenomic Data via RegLRSD Algorithm. Submitted to BMC Bioinformatics.

1.4.3 Chapter 4: Dysbiosis Index

Clinical signs of patients with IBD are highly variable, prone to inter-individual assessment, and require a significant amount of time to be gathered. These factors impede

the accurate evaluation of indices that rely on the clinical parameters such as CCECAI and CIBDAI indices. In addition to the considerable turn around time until receiving the sequencing results, sequencing-based indices (ex., MD-index and GA-map) exhibit relatively high costs. To overcome these limitations, a simple and reliable scoring system for evaluating canine IBD activity was developed. In particular, a dysbiosis index using a quantitative polymerase chain reaction (qPCR) panel composed of only 8 bacterial groups using fecal samples was built. The proposed index utilized the fact that identifying the relative abundance levels of microbes in a sample based on the PCR technology is of many orders of magnitude cheaper and faster compared to sequencing-based techniques. The results demonstrate that the DI enables discriminating healthy samples from diseased samples and tracking the response of diseased subjects to therapies with high accuracy. Additionally, the DI accurately identifies the normal microbial state, which is crucial in the fecal transportation process where it is required to identify that the donor presents balanced microbial populations. The contributions of this chapter is based on the work represented by the following papers:

- **Alshawaqfeh M**, Wajid B, Markel Guard M, Minamoto Y, Lidbury JA, Steiner JM, Serpedin E, Suchodolski JS. A Dysbiosis Index to Assess Microbial Changes in Fecal Samples of Dogs with Chronic Enteropathy. Proceedings 2016 Forum of the American College of Veterinary Internal Medicine. Denver, June CO 2016.

- **Alshawaqfeh M**, McNeely I, Lidbury JA, Steiner JM, Serpedin E, Suchodolski JS. Validation of a Dysbiosis Index to Assess Microbiota Changes in Fecal Samples of Dogs. Proceedings 2016 Forum of the European College of Veterinary Internal Medicine. Gotheburg, Sweden, September 2016.

- A Dysbiosis Index to Assess Microbial Changes in Fecal Samples of Dogs with Chronic Enteropathy. Under preparation.

2. MATRIX DECOMPOSITION FRAMEWORK FOR IDENTIFYING POTENTIAL METAGENOMIC BIOMARKERS*

2.1 Introduction

Reproducibility presents itself as a major concern for the discovery of predictive biomarkers from a wide range of biological data. Surprisingly, the reproducibility performance was not addressed by the existing state-of-the-art metagenomic biomarker discovery algorithms. Unfortunately, the biomarker detection problem from metagenomic data lacks a standard evaluation methodology that captures the key aspects of the true markers and provides a solid ground to compare competing algorithms. Therefore, we propose a protocol for evaluating a biomarker detection algorithm in terms of both (i) the consistency of the detected biomarkers and (ii) the classification performance. The proposed protocol was motivated by the model selection approach developed in [64] to find the optimal feature selection-classifier combination for a given dataset.

Currently, there are two general frameworks to tackle the problem of identifying potential markers from metagenomic data: (i) statistical framework, and (ii) machine learning framework. Deviating from these two conventional frameworks, chapter proposes to formulate the biomarker detection problem as a low rank-sparse (LRS) decomposition problem. The essence of our proposed method is to model the differentially and non-differentially abundant OTUs as a sparse and low-rank matrix, respectively. The reasoning behind this model lies in the fact that the majority of the microbes are irrelevant to the biological process at hand. Therefore, these irrelevant OTUs are supposed to have abundance levels that do not vary between two different phenotypes (i.e., healthy and dis-

*Part of this section is reprinted with permission from "Alshawaqfeh M, Bashaireh A, Serpedin E, Suchodolski J. Consistent metagenomic biomarker detection via robust PCA. Biology Direct. 2017 Jan 31;12(1):4." Copyright 2017 by the authors

eased). Hence, it is natural to consider their abundance level matrix as a low-rank matrix (denoted by $\mathbf{L}$). On the other hand, the abundance levels of the few relevant OTUs exhibit significant variations between the two phenotypes. This can be represented by a sparse matrix (denoted by $\mathbf{S}$).

The LRS decomposition framework translates the biomarker detection into the conventional robust PCA problem. Hence, the RPCA is employed to decompose the OTUs abundances matrix into the superposition of $\mathbf{L}$ and $\mathbf{S}$. Then, the bacterial biomarkers are identified based on the recovered matrix $\mathbf{S}$.

2.2 Main Contributions

The main contributions in this chapter can be summarized as follows:

- Design an evaluation protocol which provides a fair and an accurate assessment of the efficiency of a biomarker detection algorithm in terms of both (i) the consistency of the detected biomarkers and (ii) the classification performance.

- Formulate the biomarker detection problem as a low rank-sparse (LRS) matrix decomposition problem. The essence of the LRS model is to model the differentially and non-differentially abundant OTUs as a sparse and low-rank matrix, respectively. This LRS model provides several advantages. First, the multivariate nature of LRS model accounts for the complex interactions between the members of the bacterial community. This contrasts the univariate-based methods (i.e., statistical hypothesis testing and filtering techniques) that ignore such sophisticated relationships between bacteria. Second, the proposed matrix decomposition formulation is convex. This provides several benefits such as: (i) global optimality, (ii) efficient solvers, and (iii) flexibility to add convex constraints without affecting the convex structure of the problem. Third, unlike feature transformation-based algorithms, the output of the LRS decomposition is easily interpretable in the sense that it keeps the features in

their original domain.

- Propose the robust principal component analysis (RPCA) to efficiently solve this decomposition problem.

2.3 Material and Methods

2.3.1 Evaluation Protocol

A major bottleneck for the evaluation of biomarker discovery algorithms is the lack of knowledge of the true biomarkers. This hampers the objective assessment of the performance of the competing biomarker selection algorithms. To overcome this challenge, evaluation criteria have to be suitably designed in order to mimic comparisons as if the true markers were known. In particular, the evaluation metrics need to capture the features of the true biomarkers. True biomarkers are characterized by two properties. The first property is that the true markers enable distinguishing between different phenotypes. Commonly, this feature is measured via the classification performance of a classifier model built using only the selected biomarkers. The second feature is that the true signatures tend to be robust against the variation in the training set. This feature can be assessed through empirical estimation of the stability of the biomarker detection algorithm.

A common practice is to use *only* the classification performance as a measure of the effectiveness of a biomarker detection algorithm. In addition to ignoring the reproducibility performance, relying solely on the classification performance may be misleading for several reasons. First, the classification performance depends on factors other than the quality of the selected variables (i.e., biomarkers). In particular, the preprocessing steps and employed classifier model employed significantly impact the classification performance. Second, in the small sample size setups, the empirical estimation of classification accuracy may not reflect the true performance of a classifier.

The proposed protocol is based on measuring the consistency and the classification

performance over different variations of the original dataset. In particular, an empirical estimation of the consistency has been designed based on the idea that a stable biomarker detection algorithm should yield similar results under small variations of the dataset. This complies with the expectations of biologists that modifying the original dataset by adding or removing a few samples should not lead to a significant change in the identified biomarkers by an algorithm. Consequently, the procedure for estimating the consistency assumes the following steps. The first step is to repeatedly, for K times, subsample the original dataset $\mathbf{D} \in \Re_+^{p \times N}$ into two subsets: $\mathbf{D}_k^{train} \in \Re_+^{p \times \lceil rN \rceil}$ and $\mathbf{D}_k^{test} \in \Re_+^{p \times (N - \lceil rN \rceil)}$, where k stands for the iteration number. The second step is to apply the biomarker detection algorithm on the $\{\mathbf{D}_k^{train}\}_{k=1}^{K}$ subsets to find K sets of potential markers. The third step is to measure the pairwise similarity between the $\frac{K(K-1)}{2}$ pairs of the biomarker sets using a similarity or stability index. Then, the overall consistency (C_{avg}) of the algorithm is defined as the average of all pairwise similarities. Mathematically,

$$C_{avg} = \frac{2 \sum_{i=1}^{K} \sum_{j=i+1}^{K} SI(\mathcal{F}_i, \mathcal{F}_j)}{K(K-1)}, \tag{2.1}$$

where $\mathcal{F}_i$ denotes the output of the biomarker detection method over the $i'th$ subsample. $SI(\mathcal{F}_i, \mathcal{F}_j)$ represents the similarity between two marker sets measured by the stability (i.e., similarity) index SI.

Similarly, we use the same subsamples to evaluate the classification performance. Particularly, the data corresponding to the selected markers in each generated training and testing subsets are extracted and are denoted by $\mathbf{D}_k^{train}(\mathcal{F}_k)$ and $\mathbf{D}_k^{test}(\mathcal{F}_k)$, respectively. The $\mathbf{D}_k^{train}(\mathcal{F}_k)$ subset is utilized to train the classifier, while the $\mathbf{D}_k^{test}(\mathcal{F}_k)$ serves as an independent set for testing the classifier. Repeating the evaluation for K times reduces the risk of over-optimistic results for the conventional cross-validation on small-sample studies [65]. This consistency-classification evaluation protocol is summarized in Fig. 2.1.

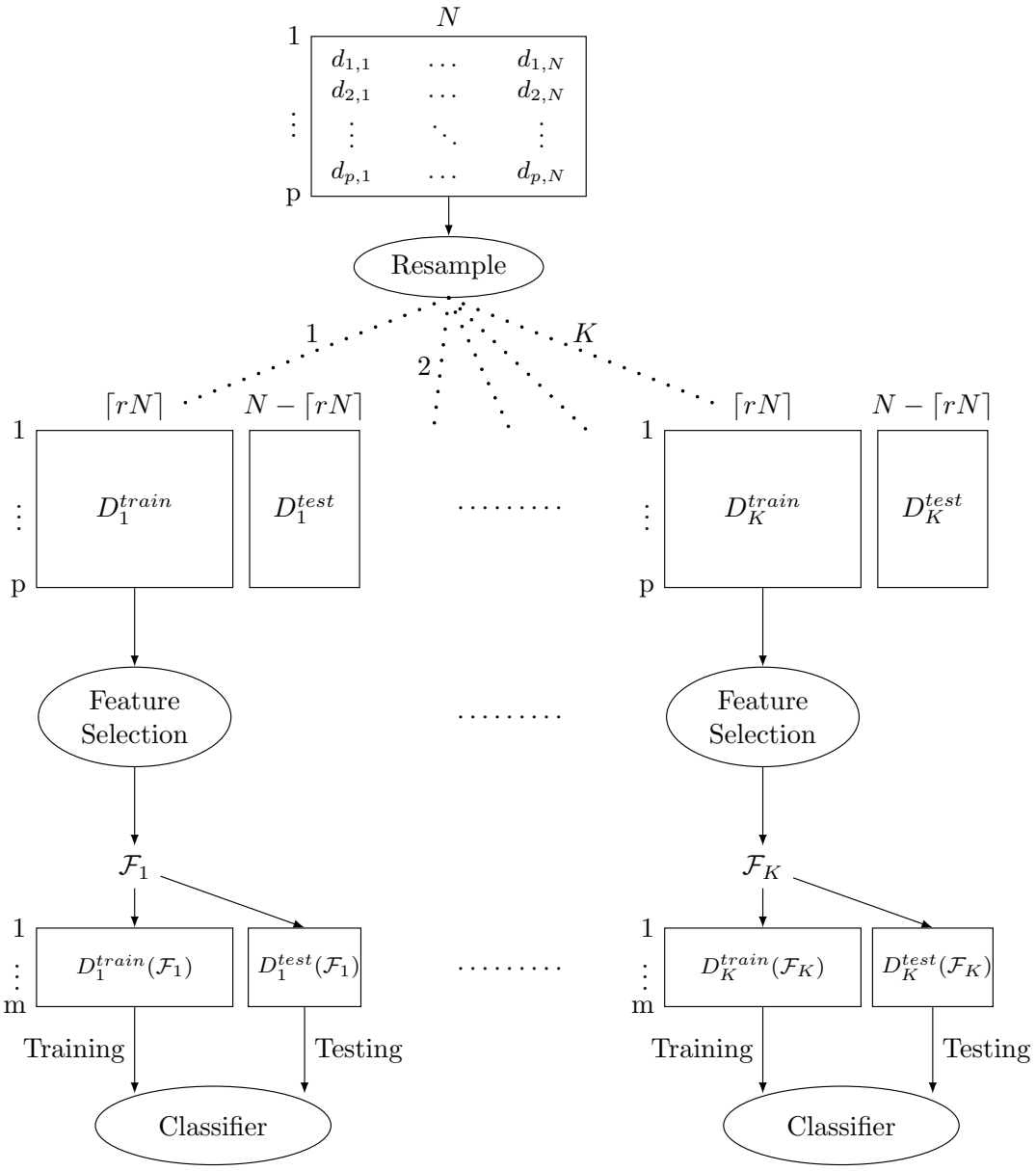

Figure 2.1: Consistency-classification evaluation protocol.

2.3.1.1 Consistency Performance

Several measures have been proposed to measure the similarity between two sets (i.e., the output of a biomarker detection algorithm over two subsamples). In this work, we

adopt the Kuncheva index (KI) [66] as a measure of similarity. KI is defined as

$$KI(\mathcal{F}_i, \mathcal{F}_j) = \frac{p.|\mathcal{F}_i \cap \mathcal{F}_j| - T^2}{T.(p-T)} = \frac{|\mathcal{F}_i \cap \mathcal{F}_j| - (T^2/p)}{T - (T^2/p)}, \tag{2.2}$$

where $T = |\mathcal{F}_i| = |\mathcal{F}_j|$. The Kuncheva index ranges from -1 to 1. The larger the value, the more common biomarkers among the two sets $\mathcal{F}_i$ and $\mathcal{F}_j$. Negative values indicate that the shared biomarkers are mostly due to chance. Negative values can be obtained due to the correction term (T^2/p) that aims to compensate for possible bias due to the randomly selected biomarkers and are common among the two marker lists.

2.3.1.2 Classification Performance

The classification performance is measured in terms of sensitivity, specificity, and accuracy. The accuracy represents the portion of the correctly classified instances in both classes (ex., healthy and diseased). In the case of imbalanced class distribution, accuracy becomes misleading since it is dominated by the majority class. This is particularly true when the prediction of the minority group is critical. Therefore, to complete the picture about the classification performance, class-specific measures such as sensitivity and specificity are also important. Sensitivity and specificity are defined as the portion of correctly predicted instances in the positive (i.e., diseased) and negative (i.e., healthy) classes, respectively. Let TN and TP denote the number of correctly identified negative and positive samples, respectively. Also, let FN and FP represent the number of false-classified samples in the negative and positive classes, respectively. Then, the accuracy, sensitivity and

specificity are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}, \tag{2.3}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{2.4}$$

$$Specificity = \frac{TN}{TN + FP}. \tag{2.5}$$

2.3.2 Low Rank-Sparse Model of Metagenomic Data

Consider the matrix $\mathbf{D} \in \Re^{p \times n}$ of bacterial abundance data, each row of $\mathbf{D}$ represents the relative abundance of an OTU in all the n samples, and each column stands for the abundance levels of all the p OTUs in one sample. In general, $p \gg n$. Therefore, it is a classical high dimension-small sample size problem. As mentioned in the Introduction section, the essence of our proposed method is to model the differentially and non-differentially abundant OTUs via a sparse and low-rank matrix, respectively. In particular, the majority of the bacterial groups are irrelevant to the biological process at hand. Therefore, these irrelevant OTUs are supposed to have abundance levels that do not vary between two different phenotypes (i.e., healthy and diseased). Hence, it is natural to consider their abundance level matrix as a low-rank matrix (denoted by $\mathbf{L}$). On the other hand, the abundance levels of the few relevant OTUs exhibit significant variations between the two phenotypes. This can be represented by a sparse matrix (denoted by $\mathbf{S}$). Mathematically,

$$\mathbf{D} = \mathbf{L} + \mathbf{S}. \tag{2.6}$$

2.3.3 Robust Principal Component Analysis

RPCA is a matrix recovery problem which aims to recover the low-rank matrix $\mathbf{L}$ and the sparse matrix $\mathbf{S}$ from their superposition $\mathbf{D}$. The authors in [67, 68] have shown that

under broad assumptions, it is possible to *exactly* recover both components (i.e., low rank and sparse matrices) by solving a convex optimization problem called *Principal Component Pursuit (PCP)*. PCP aims to minimize a weighted sum of the nuclear norm of the low-rank matrix and of the l_1 norm of the sparse matrix. Mathematically, PCP is expressed as

$$\begin{aligned} &\text{minimize } \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \\ &\text{subject to } \mathbf{D} = \mathbf{L} + \mathbf{S}, \end{aligned} \tag{2.7}$$

where λ is a positive regularization parameter that controls the sparseness and smoothness of $\mathbf{S}$ and $\mathbf{L}$, respectively. $\|\mathbf{L}\|_*$ denotes the nuclear norm of the matrix $\mathbf{L}$ and it is equal to the sum of the singular values of the matrix. $\|\mathbf{S}\|_1$ represents the l_1 norm of the matrix and it is equal to the sum of the absolute values of all the matrix entries.

Various methods have been proposed for solving the PCP problem such as the iterative thresholding approach [69] and the accelerated proximal gradient approach [70]. In this study, we adopt the augmented Lagrange multiplier (ALM) algorithm to solve (2.7). In general, ALM algorithms solve constrained optimization problems by converting them into unconstrained problems with a new objective called the *augmented Lagrangian*. The *augmented Lagrangian* for the PCP problem is given by

$$\begin{aligned} \mathcal{L}(\mathbf{L}, \mathbf{S}, \mathbf{Y}) = \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \\ \langle \mathbf{Y}, \mathbf{D} - \mathbf{L} - \mathbf{S} \rangle + \frac{\mu}{2}\|\mathbf{D} - \mathbf{L} - \mathbf{S}\|_F^2, \end{aligned} \tag{2.8}$$

where $\mathbf{Y}$ represents the Lagrange multiplier matrix, and μ stands for the single regularization parameter associated with the ALM formulation. Thus, the ALM formulation of the

PCP problem is given by

$$\begin{aligned} \text{minimize } \ \mathcal{L}(\mathbf{L},\mathbf{S},\mathbf{Y}) &= \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \\ &\langle \mathbf{Y}, \mathbf{D}-\mathbf{L}-\mathbf{S}\rangle + \frac{\mu}{2}\|\mathbf{D}-\mathbf{L}-\mathbf{S}\|_F^2. \end{aligned} \tag{2.9}$$

A standard approach to solve (2.9) is of iterative-based nature. Each iteration k consists of two steps. The first step is to solve the following sub-problem

$$(\mathbf{L}_k^*, \mathbf{S}_k^*) = \arg\min_{\mathbf{L},\mathbf{S}} \ \mathcal{L}(\mathbf{L},\mathbf{S},\mathbf{Y}_k). \tag{2.10}$$

The second step is to update the Lagrange multiplier matrix using the following equation

$$\mathbf{Y}_{k+1} = \mathbf{Y}_k + \mu\left(\mathbf{D}-\mathbf{L}_k-\mathbf{S}_k\right). \tag{2.11}$$

Since a jointly optimal solution for the sub-problem (2.10) is not available, a practical and efficient solution is to employ the alternating optimization algorithm. This alternating-based method first minimizes $\mathcal{L}(\mathbf{L},\mathbf{S},\mathbf{Y}_k)$ with respect to $\mathbf{L}$ ($\mathbf{S}$ is fixed), then it minimizes $\mathcal{L}(\mathbf{L},\mathbf{S},\mathbf{Y}_k)$ with respect to $\mathbf{S}$ ($\mathbf{L}$ is fixed). This strategy utilizes the fact that both $min_{\ \mathbf{L}}\{\mathcal{L}(\mathbf{L},\mathbf{S},\mathbf{Y})\}$ and $min_{\ \mathbf{S}}\{\mathcal{L}(\mathbf{L},\mathbf{S},\mathbf{Y})\}$ have a closed form solution. In particular, let $\mathcal{S}_\tau : \Re \to \Re$ be the shrinkage operator defined by

$$\mathcal{S}_\tau(x) = \text{sgn}(x)\text{max}(|x|-\tau, 0), \tag{2.12}$$

where $\tau \geq 0$ represents the threshold value. This shrinkage operator is extended to matri-

ces by applying it to their elements. Then,

$$\begin{aligned} \mathbf{S}^* &= \arg\min_{\mathbf{S}} \ \mathcal{L}(\mathbf{L}, \mathbf{S}, \mathbf{Y}) \\ &= \mathcal{S}_{\lambda\mu^{-1}}(\mathbf{D} - \mathbf{L} + \mu^{-1}\mathbf{Y}). \end{aligned} \tag{2.13}$$

To solve for $\mathbf{L}$, let $\mathcal{D}_\tau$ denote the singular value thresholding operator given by

$$\mathcal{D}_\tau(\mathbf{X}) = \mathbf{U}\mathcal{S}_\tau(\Sigma)\mathbf{V}^T, \tag{2.14}$$

where $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ is the singular value decomposition (SVD) of $\mathbf{X}$. Then,

$$\begin{aligned} \mathbf{L}^* &= \arg\min_{\mathbf{L}} \ \mathcal{L}(\mathbf{L}, \mathbf{S}, \mathbf{Y}) \\ &= \mathcal{D}_{\mu^{-1}}(\mathbf{D} - \mathbf{S} + \mu^{-1}\mathbf{Y}). \end{aligned} \tag{2.15}$$

Though we have a closed-form solution for $\mathbf{S}_k^*$ and $\mathbf{L}_k^*$, solving the sub-problem (2.10) requires computing (2.13) and (2.15) repeatedly until converging to the optimal solution. This repetition leads to a significant computation burden. According to [68], this burden can be avoided by updating $\mathbf{S}_k$ and $\mathbf{L}_k$ only once. Even though this does not guarantee the optimal solution of the sub-problem (2.10), it is sufficient to converge to the optimal solution of the RPCA problem as proved in [68].

2.3.4 Extracting the Differentially Abundant Bacteria via RPCA

The proposed method for identifying metagenomic biomarkers is divided into two steps. First, apply RPCA to decompose the original bacterial abundance level data into a low-rank matrix representing the non-differential abundant bacteria and a sparse matrix representing the differential abundant bacteria. Second, score each microbe (i.e., feature) by constructing a scoring vector based on the extracted sparse matrix. The top m bacteria are selected as biomarkers for the biological process under study.

As mentioned in the Introduction section, it is reasonable to consider the observed abundance matrix $\mathbf{D}$ as being the sum of a low-rank matrix $\mathbf{L}$ and a sparse matrix $\mathbf{S}$. Potential biomarkers are expected to exhibit abundance levels that vary between samples belonging to different groups. Therefore, their abundance levels can be modeled as a sparse perturbation matrix superimposed over the low-rank matrix representing the abundance levels of the non-differentiable microbes (i.e., $\mathbf{D} = \mathbf{L} + \mathbf{S}$). Consequently, the microbial biomarkers can be detected according to the sparse matrix $\mathbf{S}$. The bacteria exhibiting more variation are stronger. The extracted sparse matrix $\mathbf{S}$ can be expressed as

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pn} \end{bmatrix} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n]. \tag{2.16}$$

Corresponding to the original abundance level data matrix $\mathbf{D}$, each column contains the differential abundance levels of all the microbes in one sample, and each row of $\mathbf{S}$ represents the differential variation of a microbe in all the n samples. The entries of $\mathbf{S}$ can be either positive or negative reflecting whether the bacteria were activated or deactivated in response to the biological process. Therefore, the absolute values of the entries in $\mathbf{S}$ are needed for the identification of the differentially abundant bacteria (i.e., biomarkers). The score of the $i'th$ bacteria is calculated by summing row-wise the absolute values of the $i'th$ row in $\mathbf{S}$. Mathematically, the scoring vector ($\mathbf{v}$) is obtained by summing the absolute values of the elements of $\mathbf{S}$, and can be expressed as:

$$\mathbf{v} = \left[\sum_{j=1}^{N} |s_{1j}|, \dots, \sum_{j=1}^{N} |s_{pj}| \right]^T. \tag{2.17}$$

Large scores are associated with microbes exhibiting larger variation between the two

states. Therefore, only the genes with the top m scores are selected as biomarkers.

2.3.5 Nearest Centroid Classifier (NCC)

The nearest centroid classifier is an instance of distance-based supervised learning method. The classification process using NCC consists of two steps. The first step is to train the classifier with labeled data (i.e., $\mathbf{d}_i$) to compute the mean (i.e., centroid) of each class. The mean of the $k'th$ class (μ_{C_k}) is given by:

$$\mu_{C_k} = \frac{1}{|N_{C_k}|} \sum_{\mathbf{d}_i \in C_k} \mathbf{d}_i \tag{2.18}$$

The second step reduces to assigning a test sample ($\mathbf{z}$) to the class whose centroid is closer. Mathematically, this is equivalent to the following optimization problem:

$$\hat{C}(\mathbf{z}) = \arg\min_{C_k} dis(\mu_{C_k}, \mathbf{z}) \tag{2.19}$$

where $dis(\mu_{C_k}, \mathbf{z})$ is a distance measure between the test sample $\mathbf{z}$ and the centroid of the samples belonging to the $k'th$ class (μ_{C_k}).

2.3.6 Data Description

Unless stated otherwise, the 16S rRNA gene sequencing reads were assigned to operational taxonomic units (OTUs) using the naive Bayesian classifier employed by the Ribosomal Database Project (RDP) [71]. Reads with confidence below 80% were assigned to be uncertain. For all the datasets described below, the per-sample normalized read counts were organized in a matrix called the taxonomic relative abundances matrix. This matrix is the final input for the RPCA algorithm. As RPCA belongs to the unsupervised family of machine learning algorithms, the labels of the data are not required.

2.3.6.1 Canine Inflammatory Bowel Disease (IBD) Dataset

Naturally passed fecal samples were obtained from 89 healthy dogs and 79 dogs with chronic signs of gastrointestinal disease and confirmed inflammatory changes on histopathology. All dogs participated in different clinical studies and leftover fecal samples were utilized for this study.

Dogs with clinical signs of chronic GI disease (i.e., vomiting, diarrhea, anorexia, weight loss, etc.) were diagnosed with idiopathic IBD based on the World Small Animal Veterinary Association (WSAVA) criteria: (i) chronic (i.e., $\geq$ 3 weeks) GI signs; (ii) histopathologic evidence of mucosal inflammation; (iii) inability to document other causes of GI inflammation; (iv) inadequate response to dietary, antibiotic, and anthelmintic therapies, and (v) clinical response to anti-inflammatory or immunosuppressive agents. Histological samples were obtained endoscopically. Clinical status of each dog was evaluated using a published clinical canine IBD activity index (CIBDAI). Within the IBD dogs, 47 dogs had histological confirmed inflammation in the small intestine, 24 dogs had histological changes in both small intestine and colon, and 7 dogs had only histological changes reported in the colon. Histological changes were predominantly of lymphoplasmacytic infiltrates, with a subset of dogs also showing eosinophilic and/or neutrophilic components. Data can be downloaded from this link: https://qiita.ucsd.edu/study/description/833.

2.3.6.2 Mouse Model of Ulcerative Colitis (UC) Dataset

This dataset represents the fecal microbiota of mice model with ulcerative colitis and control mice. In particular, the microbiota of 20 T-bet$^{-/-}$ x Rag2$^{-/-}$ (UC) and 10 Rag2$^{-/-}$ (control) mice was characterized using 16S data from fecal samples. The data is publicly available in the supplementary material of [42].

2.4 Results and Discussions

This section presents the experimental evaluations on the two metagenomic studies described in the Material and Methods section. The performance of our proposed scheme, RPCA, is compared with the current state-of-the-art algorithms proposed for identifying microbial biomarkers. In particular, RPCA is compared with two statistical-based algorithms namely, MetaStats [43] and LEFSe [42], and two machine learning-based algorithms. For the machine learning-based algorithms, an entropy-based and a binary classification (BC)-based [72] filtering approach are used.

It is worth to mention that there are several implementations of the RPCA algorithm. In our experiments, we utilize the Matlab code for the exact ALM provided by the authors of [70], which is available at 'http://perception.csl.illinois.edu/matrix-rank/sample_code.html'.

The five algorithms were evaluated in terms of their classification and consistency performance according to the consistency-classification evaluation protocol shown in Fig.2.1. In our experiments, 500 subsamples (i.e., $K = 500$) were generated by randomly subsampling, without replacement, the original datasets. Due to the limited number of samples in metagenomic studies, subsamples were generated with 80% of the samples in the original dataset (i.e., $r = 0.8$). The reported results represent the average over the 500 experiments.

To reduce the dependency of the results on the classification criteria, two variants of the nearest centroid classifiers were used. In the first approach, the l_1 norm was used as a measure of distance, while in the second approach, the l_2 norm was used. In this paper, we refer to the first classifier as NCC-1 and to the second one as NCC-2. The consistency of the biomarker detection algorithms has been measured by the Kuncheva index. In order to study the impact of the number of selected features on the consistency and classification performance, the five biomarker detection algorithms were assessed at different sizes of

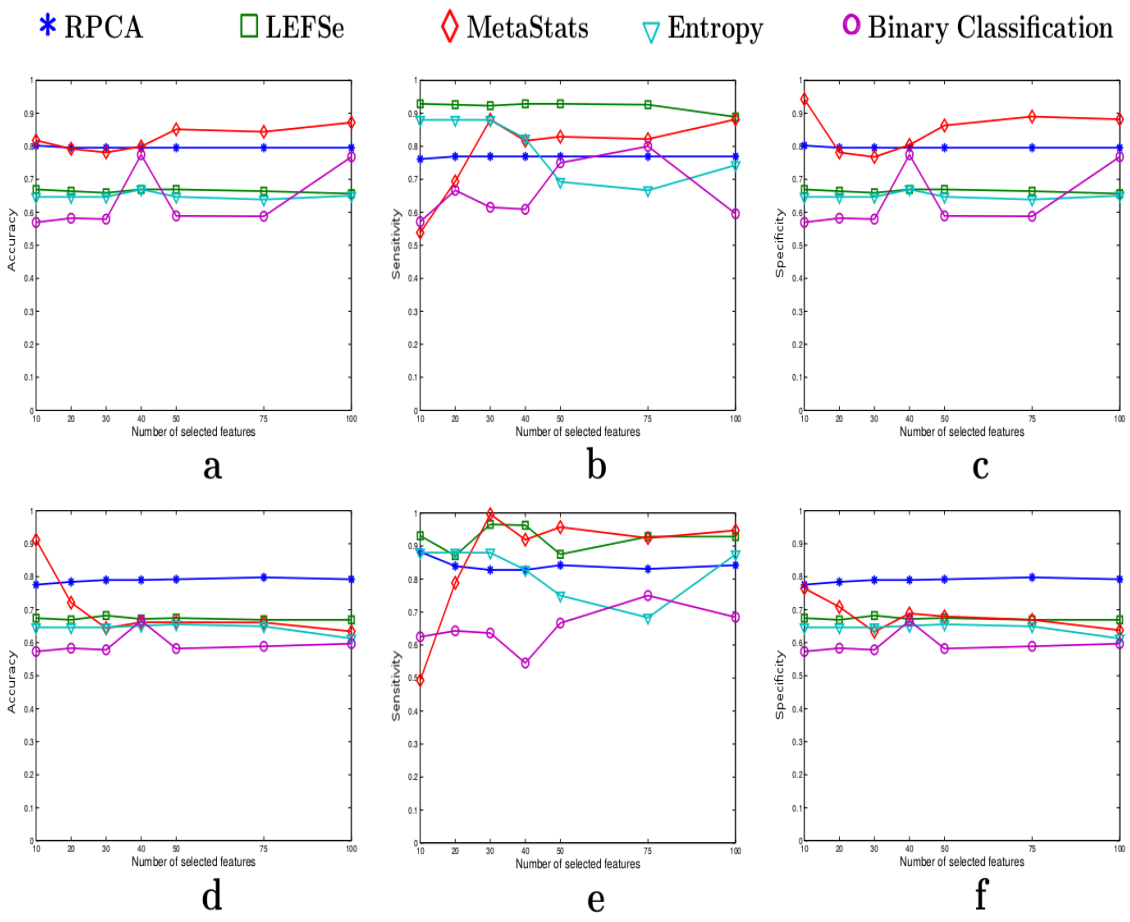

Figure 2.2: Classification performance of the five algorithms over the canine IBD dataset in terms of accuracy, sensitivity and specificity. The first row represents the results corresponding to the NCC-1 classifier (**a**: accuracy, **b**: sensitivity, **c**: specificity), while the second row represents the NCC-2 classifier results (**d**: accuracy, **e**: sensitivity, **f**: specificity).

the biomarker sets.

2.4.1 Canine Inflammatory Bowel Disease (IBD) Dataset

The performance of the five algorithms in terms of their classification accuracy for varying number of biomarkers from the canine IBD dataset is depicted in Fig. 2.2. The first row in Fig. 2.2 presents the results for the NCC-1 classifier, while the second row

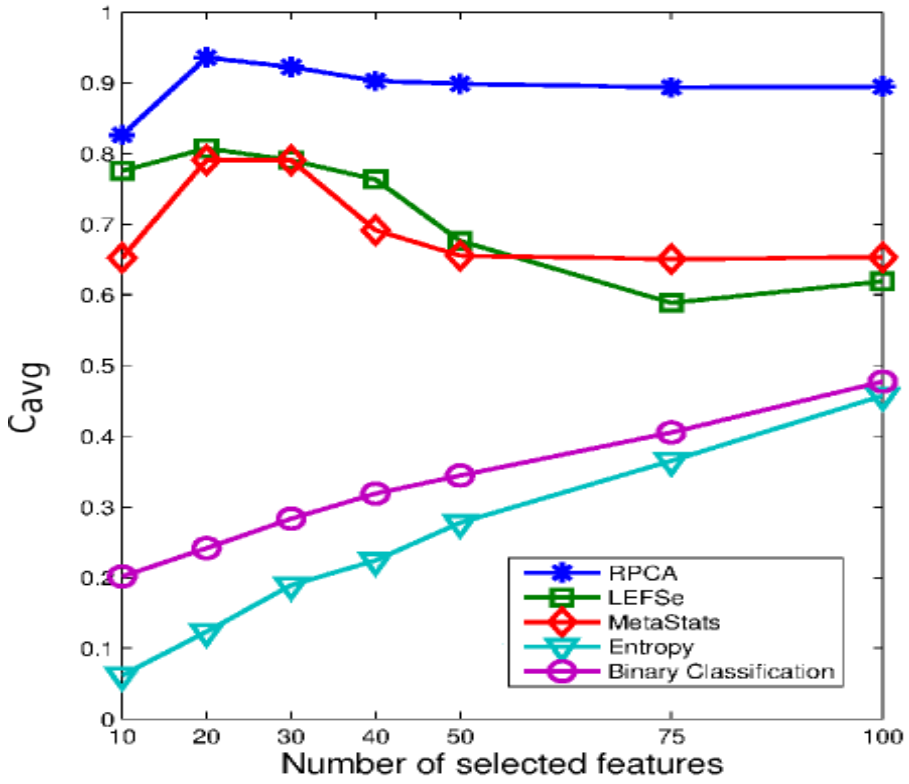

Figure 2.3: The average consistency performance measured by KI of the five biomarker discovery algorithms over the canine IBD dataset.

presents the results for the NCC-2 classifier. As the results displayed in Figs. 2.2a and b illustrate, RPCA outperforms the LEFSe, entropy-based and BC-based algorithms in terms of accuracy. In particular, the RPCA algorithm outperforms the LEFSe and entropy-based algorithms by around 13% and the BC-based algorithm by approximately 20%. The MetaStats algorithm provides comparable results to RPCA when NCC-1 is used. However, the RPBC algorithm significantly outperforms the MetaStats performance in the NCC-2 case. Moreover, RPCA provides a robust result irrespective of the variation in the applied classification method and the number of selected biomarkers. This contrasts the performance of MetaStats.

Our next simulation sought to examine the consistency performance of the five meth-

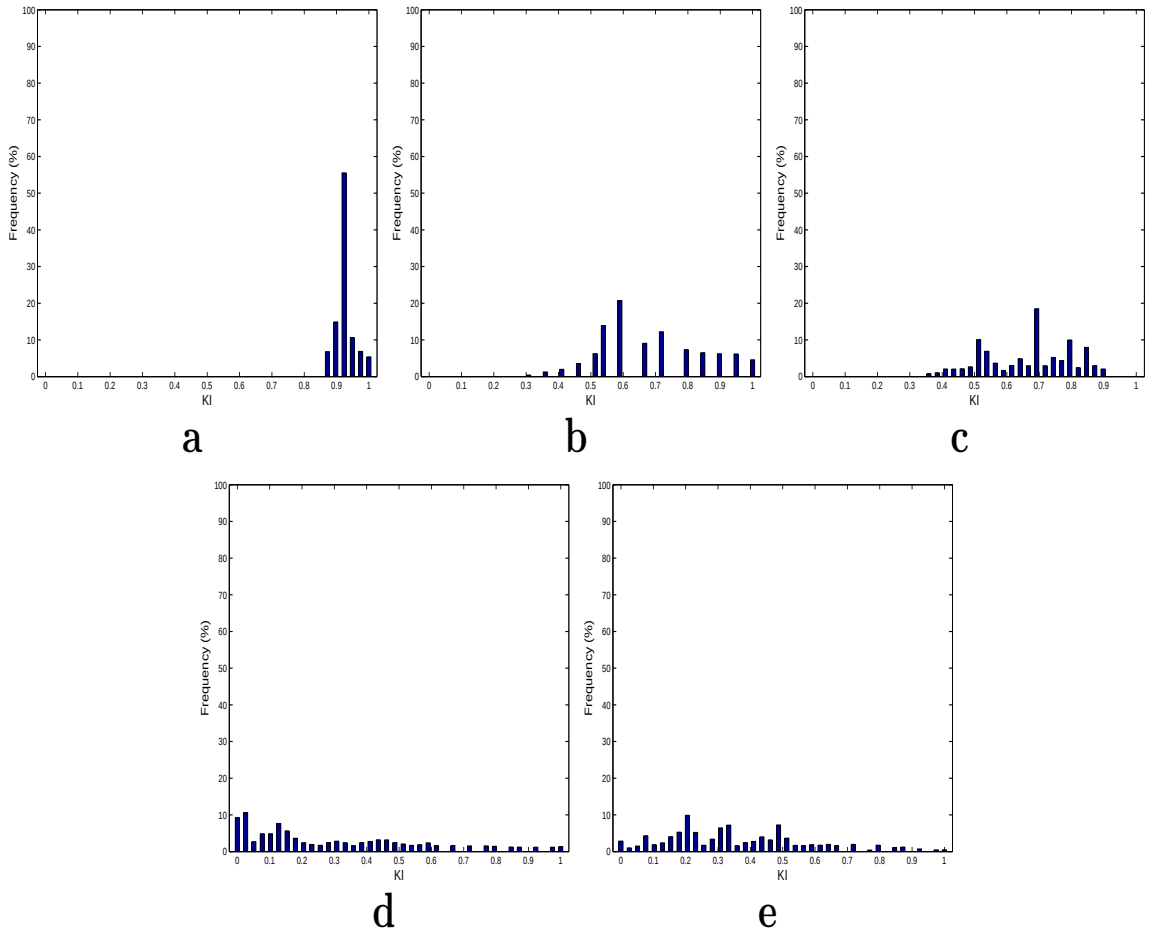

Figure 2.4: Histogram plots of all the 124750 pairwise KIs (i.e., $\frac{K(K-1)}{2} = 124750$ comparisons; $K = 500$) generated by the five biomarker discovery algorithms over the canine IBD dataset. **(a)** RPCA. **(b)** LEFSe. **(c)** MetaStats. **(d)** Entropy. **(e)** Binary Classification.

ods. Fig. 2.3 presents the KI stability values averaged over all the pairwise comparisons (i.e., $\frac{K(K-1)}{2} = 124750$ comparisons; $K = 500$). In addition to the superior consistency performance, RPCA shows a robust performance irrespective of the number of selected markers. Detailed consistency analysis when the size of the selected biomarkers equals 50 is depicted in Fig. 2.4 by presenting the histogram of the KI index computed over all pairwise comparisons. As is turns out from Fig. 2.4a, the RPCA algorithm shows a high consistent performance. This is revealed from the concentration of the histogram

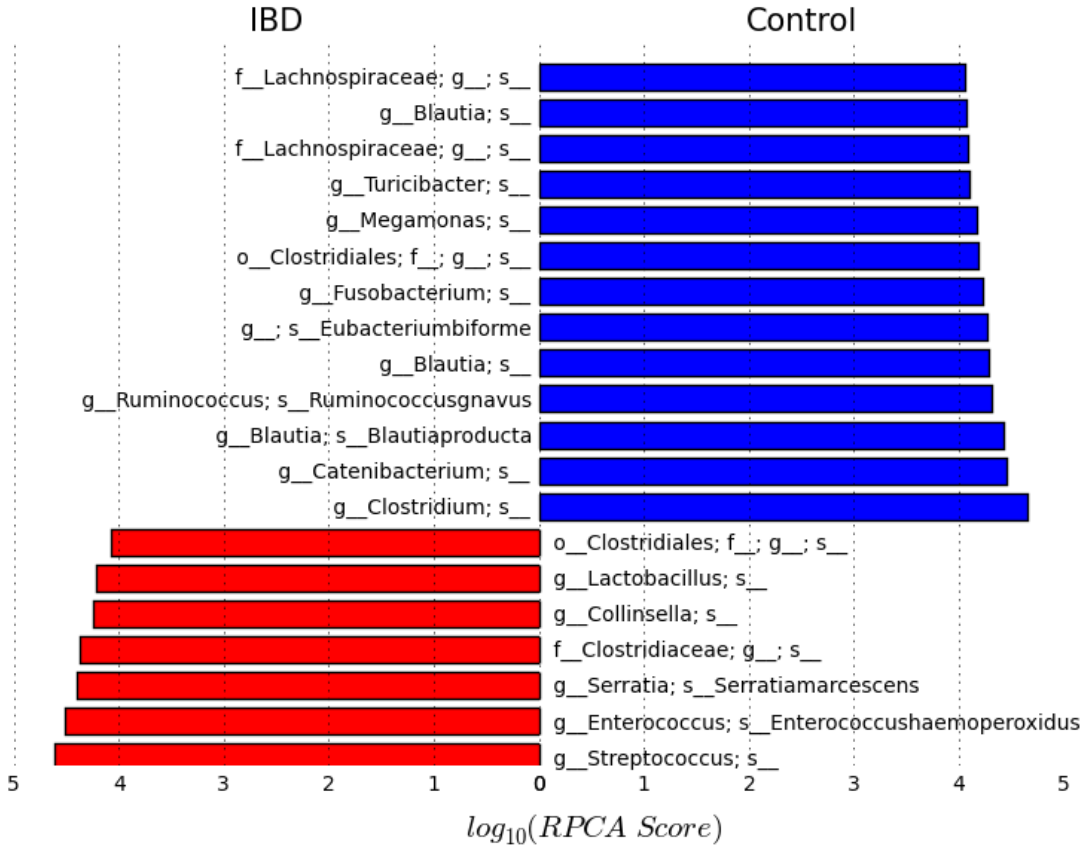

Figure 2.5: Top 20 identified biomarkers by RPCA in relation to the canine IBD dataset and their RPCA scores. Blue: the selected bacteria exhibit an increase in their abundance level in the control samples. Red: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

corresponding to RPCA at high consistency values. In particular, for almost 80% of the times, RPCA provides a stability value that is larger than or equal to 90%. On the other hand, LEFSe and MetaStats turn out to present inconsistent performance. For example, LEFSe presents KI values less than or equal to 60% for almost half of the times. The entropy-based and BC-based algorithms yield a very poor consistency performance.

The top 20 detected biomarkers by the RPCA algorithm are shown with their scores in Fig. 2.5. According to [73], Erysipelotrichaceae is considered to be a major player

in maintaining homeostasis in response to inflammation. This may explain the selection of two clades (i.e., *eubacteriumbiforme* and *g_Catenibacterium*) that belong to the Erysipelotrichaceae family as possible biomarkers for IBD. In agreement with previous studies, Collinsella [74] shows an increase in its abundance, whereas Turicibacter [75] exhibits reduced concentration in IBD subjects. This may explain selecting species belonging to these clades as potential biomarkers for IBD.

Species belong to several genera, including Blautia (i.e., Blautiaproducta and two unspecified species), Ruminococcus (i.e., Ruminococcusgnavus), and a number of taxa within the family of Lachnospiraceae show decreased abundances in IBD patients [61]. On the other hand, Lactobacillus and Streptococcus exhibit an increase in their abundance levels in patients with Crohn's disease [61]. Fusobacterium has previously been suggested as a biomarker for IBD [76]. In order to validate the detected markers by RPCA, an independent validation experiment has been conducted. In particular, quantitative PCR (qPCR) assays targeting thirteen bacterial groups were conducted over fecal DNA samples taken from 285 healthy dogs and 172 dogs with chronic enteropathy (CE) [77]. In this experiment, the final PCR panel (i.e., Faecalibacterium, Turicibacter, E. coli, Streptococcus, Blautia, and Fusobacterium) includes five OTUs (i.e., Faecalibacterium, Turicibacter, Streptococcus, Blautia, and Fusobacterium) that are strongly suggested as potential signatures for IBD by our RPCA-based algorithm.

The top 30 detected markers by the five algorithms is listed in Appendix A. As shown in Appendix A, the RPCA is the only algorithm that has suggested Faecalibacterium and Blautia as potential biomarkers for IBD. In particular, Blautia has been proposed as a strong driver for IBD by RPCA and three species from Blautia genera (rank 5, 9, and 18) were selected in the top 30 markers. Moreover, the Streptococcus has been strongly suggested as a strong potential marker for IBD by RPCA. Specifically, one species of Streptococcus was ranked second by RPCA. This agrees with Metastats which suggested

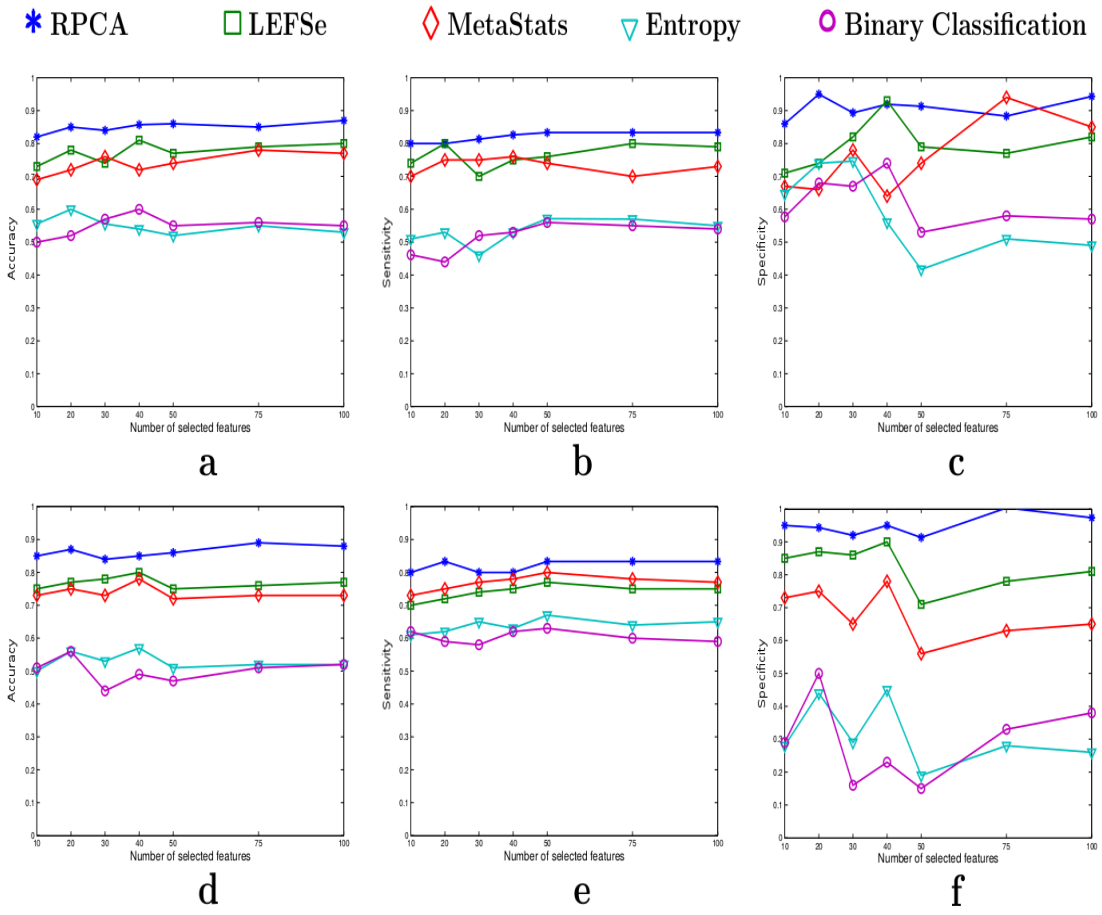

Figure 2.6: Classification performance of the five algorithms over the mouse model of UC dataset in terms of accuracy, sensitivity and specificity. The first row represents the results corresponding to the NCC-1 classifier (**a**: accuracy, **b**: sensitivity, **c**: specificity), while the second row represents the NCC-2 classifier results (**d**: accuracy, **e**: sensitivity, **f**: specificity).

two species belonging to Streptococcus genera (rank 3 and 8), and BC which suggested one Streptococcus species with rank 23 as IBD marker. On the other hand, LEFSe and Entropy algorithms do not include Streptococcus in their suggested lists of IBD markers. RPCA and LEFSe are the only algorithms that suggested Turicibacter as a signature for IBD. Fusobacterium was strongly recommended as a possible marker for IBD by RPCA and MetaStats (rank 12 and 5, respectively) while it is less favored by LEFSe (rank 25).

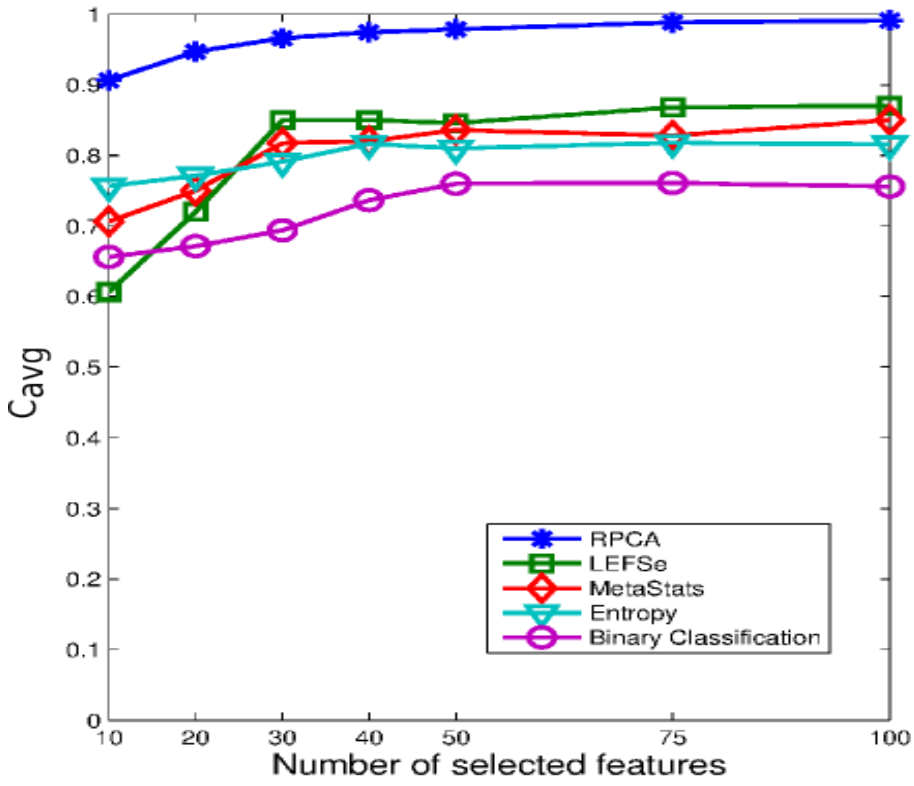

Figure 2.7: The average consistency performance measured by KI of the five biomarker discovery algorithms over the mouse model of UC dataset.

This independent validation experiment demonstrates the efficiency of the RPCA-based algorithm in identifying markers with high classification potential.

2.4.2 Mouse Model of Ulcerative Colitis (UC) Dataset

Fig. 2.6 presents the classification performance of the five algorithms for varying number of biomarkers from the ulcerative colitis mice model dataset. The first and second row represents the classification performance corresponding to the NCC-1 and NCC-2 classifier, respectively. The results in Figs. 2.6a and d demonstrate that the RPCA algorithm outperforms all the four methods in terms of classification accuracy. Moreover, RPCA exhibits a consistent performance regardless of the classification method and the number of

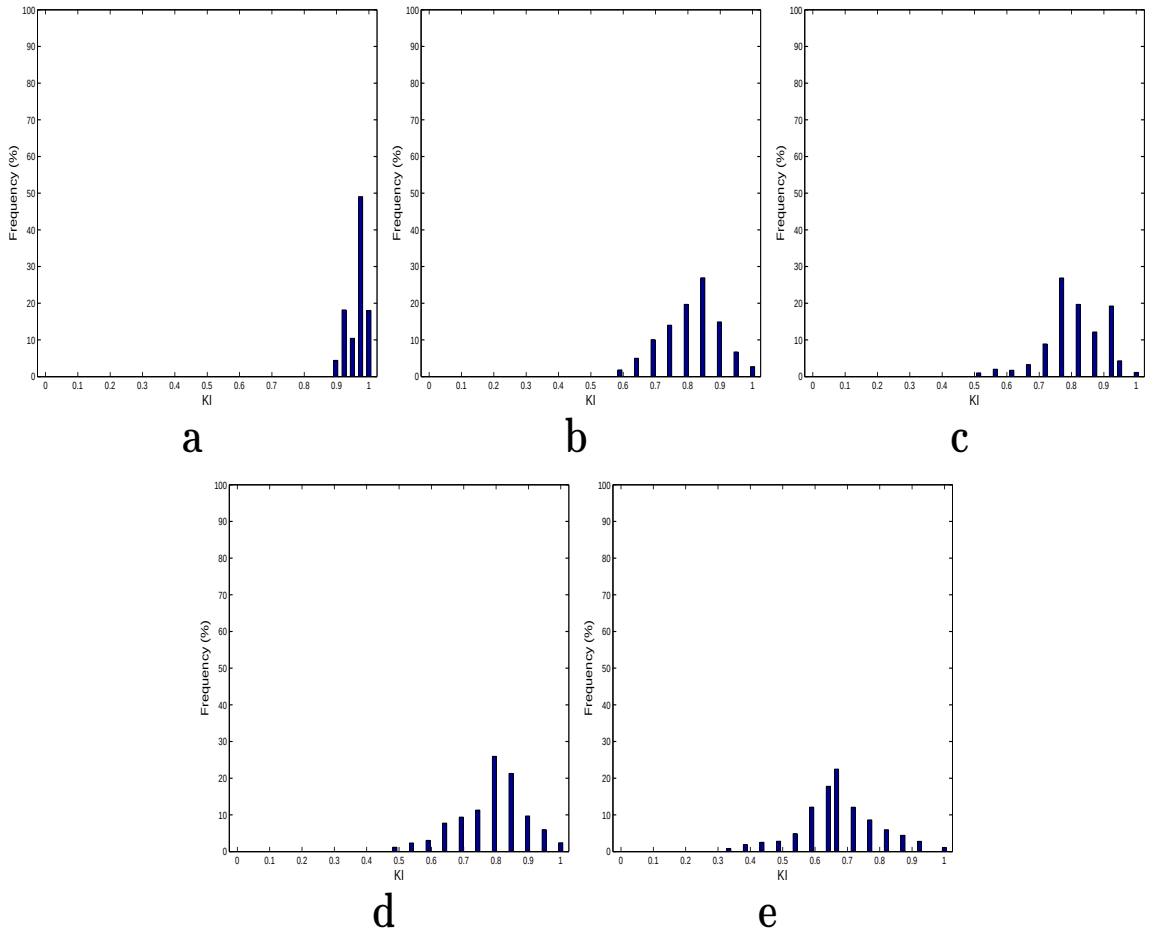

Figure 2.8: Histogram plots of all the 124750 pairwise KIs (i.e., $\frac{K(K-1)}{2} = 124750$ comparisons; $K = 500$) generated by the five biomarker discovery algorithms over the mouse model of UC dataset. **(a)** RPCA. **(b)** LEFSe. **(c)** MetaStats. **(d)** Entropy. **(e)** Binary Classification.

biomarkers included in the classifier models. On the other hand, the other four algorithms exhibit a variation in their accuracy by around 10% when varying the number of selected markers from 10 to 100.

The average KI values over all the pairwise comparisons and their histogram when the number of selected biomarkers equals 30 are depicted in Figs. 2.7 and 2.8, respectively. Fig. 2.7 points out that RPCA exhibits a very high consistency performance and

outperforms all the other algorithms. In particular, when the number of markers is larger than 30, RPCA provides an improvement by around 10-15% over the LEFSe, MetaStats and entropy-based algorithms, and approximately 20% over the BC-based method. This improvement increases when the number of markers is less than 30. For example, for a marker set of size 10, this gain increases to 30% and 20% when RPCA is compared to MetaStats and LEFSe, respectively.

The distributions of the all pairwise KI for the five algorithms are depicted in Fig. 2.8. These histograms provide a finer view of the consistency performance of the algorithms. Fig. 2.8a demonstrates that the RPCA algorithm provides a consistent performance as the KI values exceed 95% for almost 80% of the times. The other algorithms show much less consistent performance compared to RPCA. This behavior is clear from the facts that the histograms of these methods are centered at lower values for KI and the wide spread of KI values.

The top 10 identified biomarkers by the RPCA algorithm are listed in Fig. 2.9. RPCA suggests the enrichment of Oscillibacter, Alistipes, Helicobacter and Escherichia/Shigella as potential biomarkers for UC. This agrees with previous studies. For example, the authors of [78] found that Alistipes presents a very low abundance level in almost all patients diagnosed with UC. The previous study [39] reported that consistent reductions of acetate producer clades such as Ruminococcaceae, to which Oscillibacter belongs, may negatively impact the host ability to repair the epithelium and to regulate inflammation. For Helicobacter, the authors of [79] reported significantly lower rates of Helicobacter pylori, the most widely known species of Helicobacter genus, in UC patients. Also, the increased level of the Escherichia/Shigella has been linked to the intestinal inflammation [80].

In agreement with the previous studies, RPCA associates the reduction in the concentration of Lactobacillus, Bifidobacterium and Bacteroides to UC. Previous studies have reported similar results. For example, decreased concentrations of Lactobacillus and Bifi-

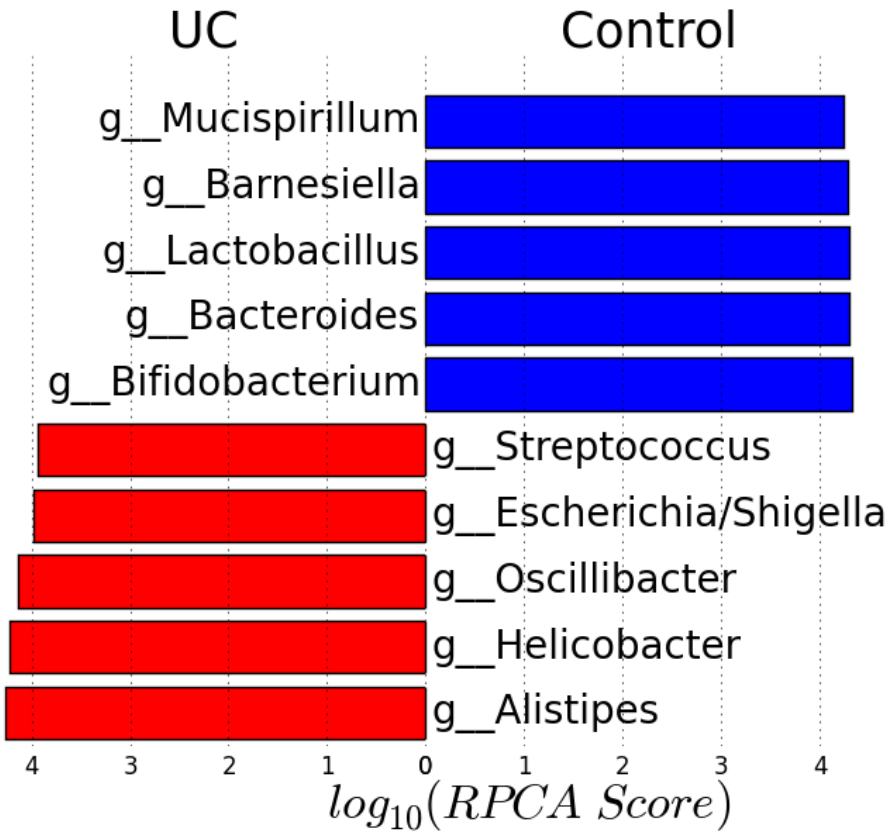

Figure 2.9: Top 10 identified biomarkers by RPCA in relation to the mouse model of ulcerative colitis dataset and their RPCA scores. Blue: the selected bacteria exhibit an increase in their abundance level in control samples. Red: the selected bacteria exhibit an increase in their abundance level in UC samples.

dobacterium in colonic biopsy specimens and reduced fecal concentrations of lactobacilli and bifidobacteria have been found in patients with active UC [81, 82]. According to [83], the UC can be characterized by the decrease in the abundance levels of Bacteroides. This analysis highlights the agreement of RPCA with the biological findings and suggests additional taxa as possible biomarkers for UC.

The top 10 detected markers by the five algorithms is listed in Appendix B. As is clear, 8 out of the 10 identified markers by RPCA and LEFSe are in common. Instead

of Desulfovibrio and Butyricicoccus which are identified by LEFSe, RPCA suggested Bacteroides and Lactobacillus as potential markers for UC. RPCA shows less agreement with MetaStats in which only 3 signatures are in common between them. Specifically, RPCA and MetaStats agrees on Bifidobacterium, Streptococcus, and Escherichia/Shigella as possible markers for UC. On the other hand, RPCA does not share any of its detected markers neither with the Entropy-based method nor with the BC algorithm.

2.5 Summary

This chapter addresses two essential challenges associated with detecting potential markers from metagenomic data. The first challenge is the lack of knowledge of the true biomarkers, which hampers the objective assessment of the performance of competing biomarker selection algorithms. The second challenge is to design a marker selection algorithm that provides reliable performance in terms of reproducibility and classification accuracy. Toward this end, an evaluation protocol that mimics comparisons as if the true markers were known was proposed. The essence of this evaluation protocol is to measure the consistency and the classification performance over different variations of the original dataset, as shown in Fig. 2.1. These variations are generated by means of random subsampling without replacement. Moreover, we proposed a RPCA-based biomarker detection algorithm for the problem of identifying possible bacterial markers. This matrix decomposition framework enables the characterization of specific microbial taxa that are differentially expressed between samples belonging to two different classes.

Comprehensive comparisons with state-of-the-art biomarker discovery algorithms belonging to the class of statistical methods and the class of machine learning approaches were conducted. The obtained results were evaluated (i) statistically in terms of classification accuracy and reproducibility performance and (ii) biologically by discussing their biological relevance to the case under study and their agreement with previous studies.

Experiments were conducted on two realistic datasets. The first dataset is in relation to healthy dogs and dogs diagnosed with IBD. The second dataset is a mouse model of ulcerative colitis. Experiments show that the RPCA algorithm effectively detects microbial biomarkers in both datasets. In particular, the detected biomarkers by the RPCA algorithm exhibit high accuracy in discriminating the metagenomic samples belonging to different phenotypes. More importantly, RPCA shows a high reproducibility performance when compared with the other algorithms. These findings demonstrate that (i) the concept of modeling the abundance level matrix as the sum of a low-rank matrix representing the irrelevant bacteria and a sparse matrix containing the abundances of informative bacteria, (ii) the use of RPCA to recover this sparse matrix, and (iii) the inherent multivariate nature of RPCA that handles the complex microbial interactions, were successful in finding potential metagenomic biomarkers with high reproducibility and discriminative power.

3. INCORPORATING PRIOR KNOWLEDGE FOR IMPROVING METAGENOMIC BIOMARKER DISCOVERY *

3.1 Introduction

Due to the increasing number of studies that links the distortion of the bacterial balance to certain host health and disease states, there has been much interest in identifying potential microbial markers from metagenomic data. The instability of a biomarker detection algorithm renders the identified markers questionable and hinders the translation of these findings into clinical application. In practice, the number of available samples varies from experiment to experiment. Therefore, a robust biomarker detection algorithm is needed to provide a set of potential markers irrespective of the number of available samples. The main reason for the instability of a biomarker detection algorithm is ignoring the stability in the design process of the algorithm [63]. To mitigate this problem, the authors suggest incorporating prior knowledge to guide the algorithm toward more accurate results. For example, assigning higher weights for relevant features yields improved performance in terms of stability and classification [84].

This chapter proposes a novel Regularized Low Rank-Sparse Decomposition (RegLRSD) algorithm. RegLRSD extends the LRS model (2.6) by incorporating the prior knowledge in the biomarker identification process. In particular, the fact that the abundance profiles of non-informative bacteria do not exhibit significant variation is utilized by adding a smoothness constraint on the recovered low-rank matrix. To solve this matrix decomposition problem, an efficient solution based on the alternating direction method of multipliers (ADMM) is proposed.

*Part of this section is from "Reliable Biomarker Discovery from Metagenomic Data via RegLRSD Algorithm." Submitted to BMC Bioinformatics.

3.2 Material and Methods

3.2.1 Extracting the Sparse Matrix via RegLRSD

Based on the low rank-sparse decomposition model of the bacterial abundance profiles (2.6), identifying potential biomarkers boils down to a matrix decomposition problem, with the aim of extracting the sparse matrix. The authors in [69, 67] showed that under broad assumptions, it is possible to *exactly* recover both components (i.e., low rank and sparse matrices) by solving a convex optimization problem, called Principal Component Pursuit (PCP).

In an attempt to improve the accuracy of estimating $\mathbf{S}$ and $\mathbf{L}$, we extend the formulation in (2.7) by adding a penalty term in order to enforce the smoothness of each row of $\mathbf{L}$. This penalty term incorporates the prior knowledge that the abundance profiles of non-differentially abundant OTUs are smooth. In this study, the first order difference (FOD) is adopted as a measure of smoothness, and it is defined as:

$$\|\mathbf{X}\|_{FOD} = \sum_{j} \|\mathbf{F}\mathbf{x}_j\|_1, \tag{3.1}$$

where $\mathbf{x}_j$ denotes the j^{th} column of $\mathbf{X}$, and $\mathbf{F}$ represents the first order difference operator defined as:

$$\mathbf{F} = \begin{bmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}. \tag{3.2}$$

Thus, the RegLRSD algorithm aims to solve the following optimization problem:

$$(\mathbf{L}^*, \mathbf{S}^*) = \arg\min_{\mathbf{L},\mathbf{S}} \{ f(\mathbf{D}, \mathbf{L}, \mathbf{S}) = \frac{1}{2}\|\mathbf{D} - \mathbf{L} - \mathbf{S}\|_F^2 + \alpha\|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \beta \sum_{i=1}^{p} \|\mathbf{F}\mathbf{l}_i^T\|_1 \}, \tag{3.3}$$

where $\mathbf{l}_i^T$ stands for the i^{th} row of $\mathbf{L}$. Variables α and β are regularization parameters. One key advantage of this formulation is that that the optimization problem (3.3) is convex. This convex formulation provides the following benefits: (i) it ensures a globally optimal solution, (ii) it enables utilizing the well-established theory and tools for solving convex problems, and (iii) it offers the flexibility of adding additional convex constraints to reflect additional prior knowledge. However, direct application of generic convex solvers may not be feasible due to the high dimension nature of our problem. For example, interior point methods exhibit high-order complexity. Moreover, a jointly optimal solution for the optimization problem (3.3) is not available. Therefore, we propose an efficient alternating-based algorithm to solve (3.3). This alternating-based method first minimizes $f(\mathbf{L}, \mathbf{S})$ with respect to $\mathbf{S}$ ($\mathbf{L}$ is fixed), then it minimizes $f(\mathbf{L}, \mathbf{S})$ with respect to $\mathbf{L}$ ($\mathbf{S}$ is fixed). In particular, it adopts the following updating steps:

$$\mathbf{S}^{(k)} = \arg\min_{\mathbf{S}} \; f(\mathbf{L}^{(k-1)}, \mathbf{S}) \tag{3.4}$$

$$\mathbf{L}^{(k)} = \arg\min_{\mathbf{L}} \; f(\mathbf{L}, \mathbf{S}^{(k)}). \tag{3.5}$$

This strategy utilizes the fact that the two sub-problems (3.4) and (3.5) admit efficient solutions. In particular, the problem in (3.4) can be rewritten as follows:

$$\mathbf{S}^{(k)} = \arg\min_{\mathbf{S}} \; \frac{1}{2}\|\mathbf{D} - \mathbf{L}^{(k-1)} - \mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_1. \tag{3.6}$$

Problem (3.6) admits the following closed form solution:

$$\mathbf{S}^{(k)} = \mathcal{S}_{\lambda}(\mathbf{D} - \mathbf{L}^{(k-1)}), \tag{3.7}$$

where $\mathcal{S}_{\tau} : \Re \rightarrow \Re$ denotes the *shrinkage operator* defined by:

$$\mathcal{S}_{\tau}(x) = \text{sgn}(x)\text{max}(|x| - \tau, 0), \tag{3.8}$$

where $\tau \geq 0$ represents the threshold value. This shrinkage operator is extended to a matrix by applying it to its elements. The problem in (3.5) can be rewritten as:

$$\begin{aligned} \mathbf{L}^{(k)} = \arg\min_{\mathbf{L}} \ & \frac{1}{2}\|\mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L}\|_F^2 + \alpha\|\mathbf{L}\|_* \\ & + \beta \sum_{i=1}^{p} \|\mathbf{Fl}_i^T\|_1. \end{aligned} \tag{3.9}$$

The current formulation of the optimization problem in (3.9) is neither in a format that admits a closed form solution as (3.4) nor in the format of a well-established problem that admits an efficient solution. Moreover, relying on generic convex techniques to solve (3.9) may not be efficient. The difficulty in this minimization problem arises from the combination of the two non-smooth terms $\|\mathbf{L}\|_*$ and $\sum_{i=1}^{p} \|\mathbf{Fl}_i^T\|_1$. Therefore, we propose to reformulate (3.9) by introducing an additional variable and constraint to separate these two terms. Adding this auxiliary variable enables the decomposition of (3.9) into two subproblems that can be solved efficiently. The first subproblem is the *nuclear-norm regularized least-squares* problem which admits a closed form solution [85]. The second problem can be recast as the *total variation denoising* problem [86], which presents an efficient solution [87]. In particular, (3.9) is reformulated as:

$$
\begin{aligned}
(\mathbf{L}, \mathbf{Y}) = &\arg\min_{\mathbf{L},\mathbf{Y}} \ \frac{1}{2}\|\mathbf{D} - \mathbf{S}^{(k)} + \mathbf{L}\|_F^2 + \alpha\|\mathbf{L}\|_* \\
&+ \beta \sum_{i=1}^{p} \|\mathbf{F}\mathbf{y}_i^T\|_1, \\
&\text{subject to } \ \mathbf{Y} = \mathbf{L},
\end{aligned} \tag{3.10}
$$

where $\mathbf{y}_i^T$ stands for the i^{th} row of the auxiliary variable $\mathbf{Y}$. To solve (3.10), we employ the alternating direction method of multipliers (ADMM) [87]. In general, the ADMM algorithm converts the constrained optimization problem into unconstrained problems with a new objective called the augmented Lagrangian. The augmented Lagrangian associated with the optimization problem (3.10) is given by:

$$
\begin{aligned}
\mathcal{L}_\rho(\mathbf{L}, \mathbf{Y}, \mathbf{Z}) = &\frac{1}{2}\|\mathbf{D} - \mathbf{S}^{(k)} + \mathbf{L}\|_F^2 + \alpha\|\mathbf{L}\|_* \\
&+ \beta \sum_{i=1}^{p} \|\mathbf{F}\mathbf{y}_i^T\|_1 + \langle \mathbf{Z}, \mathbf{L} - \mathbf{Y} \rangle \\
&+ \frac{\rho}{2}\|\mathbf{L} - \mathbf{Y}\|_F^2,
\end{aligned} \tag{3.11}
$$

where $\mathbf{Z}$ represents the Lagrange multiplier matrix. Thus, the ADMM formulation of (3.10) is given by:

$$
(\mathbf{L}, \mathbf{Y}, \mathbf{Z}) = \arg\min_{\mathbf{L},\mathbf{Y},\mathbf{Z}} \ \mathcal{L}_\rho(\mathbf{L}, \mathbf{Y}, \mathbf{Z}). \tag{3.12}
$$

The ADMM solution of (3.12) is of iterative-based nature. Each iteration r consists of the following update steps:

$$
\begin{aligned}
\mathbf{L}^{(r)} = &\arg\min_{\mathbf{L}} \ \frac{1}{2}\|\mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L}\|_F^2 \\
&+ \alpha\|\mathbf{L}\|_* + \left\langle \mathbf{Z}^{(r-1)}, \mathbf{L} - \mathbf{Y}^{(r-1)} \right\rangle \\
&+ \frac{\rho}{2}\|\mathbf{L} - \mathbf{Y}^{(r-1)}\|_F^2,
\end{aligned} \tag{3.13}
$$

$$\begin{aligned}\mathbf{Y}^{(r)} = \arg\min_{\mathbf{Y}} \; & \left\langle \mathbf{Z}^{(r-1)}, \mathbf{L}^{(r)} - \mathbf{Y} \right\rangle \\ & + \frac{\rho}{2}\|\mathbf{L}^{(r)} - \mathbf{Y}\|_F^2 + \beta \sum_{i=1}^{p} \|\mathbf{F}\mathbf{y}_i^T\|_1,\end{aligned} \tag{3.14}$$

$$\mathbf{Z}^{(r)} = \mathbf{Z}^{(r-1)} + \rho(\mathbf{L}^{(r)} - \mathbf{Y}^{(r)}) \tag{3.15}$$

Remark-1: For any arbitrary vectors $\mathbf{u}, \mathbf{v} \in \Re^n$, and scalars $a, b \in \Re$, the following relation holds:

$$\langle a\mathbf{v} + b\mathbf{u}, \mathbf{u} \rangle = b \left\| -\frac{a}{2b}\mathbf{v} - \mathbf{u} \right\|_F^2 - \frac{a^2}{4b} \|\mathbf{v}\|_F^2 . \tag{3.16}$$

The proof of (3.16) is provided in Appendix C.1. Based on Remark-1, the problem in (3.13) can be recast as:

$$\begin{aligned}\mathbf{L}^{(r)} = \arg\min_{\mathbf{L}} \; & \alpha\|\mathbf{L}\|_* + \\ & \frac{1+\rho}{2} \left\| \frac{\mathbf{D} - \mathbf{S}^{(k)} + \rho\mathbf{Y}^{(r-1)} - \mathbf{Z}^{(r-1)}}{1+\rho} - \mathbf{L} \right\|_F^2\end{aligned} \tag{3.17}$$

Detailed derivation of (3.17) is given in Appendix C.2. According to [85], problem (3.17) admits the following closed form solution:

$$\mathbf{L}^{(r)} = \mathcal{D}_{\frac{\alpha}{1+\rho}} \left(\frac{\mathbf{D} - \mathbf{S}^{(k)} + \rho\mathbf{Y}^{(r-1)} - \mathbf{Z}^{(r-1)}}{1+\rho} \right), \tag{3.18}$$

where $\mathcal{D}_\tau$ is the *singular value shrinkage* operator defined by:

$$\mathcal{D}_\tau(\mathbf{X}) = \mathbf{U}\mathcal{D}_\tau(\Sigma)\mathbf{V}^T, \quad \mathcal{D}_\tau(\Sigma) = \text{diag}(\{\sigma_i - \tau\}_+) \tag{3.19}$$

where $\mathbf{U}$, $\mathbf{V}$, and σ_i represent the left singular vectors, the right singular vectors and the singular values of the matrix $\mathbf{X}$, respectively, and $(x)_+$ stands for the positive part of x (i.e., $(x)_+ = \max(0, x)$). In other words, $D_\tau(\mathbf{X})$ applies a soft-thresholding rule to the singular values of $\mathbf{X}$, effectively shrinking these towards zero. This is the reason why this

transformation is referred to as the *singular value shrinkage* operator. Following the same derivation lines of (3.17) (see Appendix C.2), problem (3.14) can be recast as:

$$\begin{aligned}\mathbf{Y}^{(r)} = \arg\min_{\mathbf{Y}} \ & \frac{\rho}{2}\left\|\frac{\mathbf{Z}^{(r-1)}+\rho\mathbf{L}^{(r)}}{\rho}-\mathbf{Y}\right\|_F^2 \\ & +\beta\sum_{i=1}^{p}\|\mathbf{F}\mathbf{y}_i^T\|_1.\end{aligned} \tag{3.20}$$

The rows of $\mathbf{Y}$ can be updated separately according to the following optimization problem:

$$\begin{aligned}\mathbf{y}_i^{T(r)} = \arg\min_{\mathbf{y}} \ & \frac{\rho}{2}\left\|\frac{\mathbf{z}_i^{T(r-1)}+\rho\mathbf{l}_i^{T(r)}}{\rho}-\mathbf{y}\right\|_F^2 \\ & +\beta\|\mathbf{F}\mathbf{y}\|_1,\end{aligned} \tag{3.21}$$

where $\mathbf{z}_i$ and $\mathbf{l}_i$ are the i^{th} rows of $\mathbf{Z}$ and $\mathbf{L}$, respectively. Problem (3.21) is often called the total variation denoising problem [86], and it admits an efficient solution via ADMM as described in Section 6.4.1 in [87].

The proposed RegLRSD algorithm is summarized in Algorithm 1.

Algorithm 1: RegLRSD algorithm to solve the regularized low rank-sparse matrix decomposition problem (3.3).

```
Input  : D
while not convereged do
    update S^k using equation (3.7);
    while not convereged do
        update L^r using equation (3.18);
        update Y^r by solving (3.21) using ADMM solver;
        update Z^r using equation (3.15);
    end
    L^k ← L^r;
end
Output: L, S
```

3.2.2 Extracting the Differentially Abundant Bacteria via RegLRSD

The proposed method for biomarkers discovery consists of two phases. First, apply RegLRSD to decompose the original bacterial abundance level data into a low-rank matrix representing the non-differential abundant bacteria and a sparse matrix representing the differential abundant bacteria. Second, construct a scoring vector based on the extracted sparse matrix to rank each OTU (i.e., feature). Then, the m highest scores OTUs are selected as potential bacterial biomarkers for the biological process under study.

The reasoning for employing the sparse matrix for extracting the potential biomarkers is that the abundance levels of informative OTUs can be modeled as a sparse perturbation matrix superimposed over the low-rank matrix representing the abundance levels of the non-informative microbes (i.e., $\mathbf{D} = \mathbf{L} + \mathbf{S}$). The stronger the variation in the abundance levels of OTUs, the larger the magnitude of the corresponding elements in the sparse matrix $\mathbf{S}$. It is pertinent to mention that the strength of the variation of each OTU between two phenotypes is determined by the absolute values of the non-zero elements in $\mathbf{S}$ rather than their exact values. This is because the entries of $\mathbf{S}$ can be either positive or negative based on the role (i.e., activation or deactivation) of the corresponding microbes in the biological process. Therefore, the score of the i^{th} OTU is obtained by summing the absolute values of the i^{th} row in $\mathbf{S}$. Mathematically, the scoring vector $\mathbf{v}$ is given by:

$$\mathbf{v} = \left[\sum_{j=1}^{n} |s_{1j}|, \ldots, \sum_{j=1}^{n} |s_{pj}| \right]^T . \tag{3.22}$$

3.2.3 Parameter Selection

RegLRSD algorithm has four regularization parameters, α, β, λ and ρ that control the impact of the rank (i.e., $\|\mathbf{L}\|_*$), smoothness (i.e., $\sum_{i=1}^{p} \|\mathbf{Fl}_i^T\|_1$), sparseness (i.e., $\|\mathbf{S}\|_0$), and fitness (i.e., $\|\mathbf{L} - \mathbf{Y}\|_F^2$) penalties in (3.3) and (3.11). In order to select values for these

parameters, we relied on similar models and utilized the recommended settings proposed in literature. For example, the PCP problem (2.7), which is a simplified version of the objective of RegLSRD algorithm, was addressed in [67]. In particular, PCP assumes the following objective $\|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_0$. The authors in [67] proved that under mild assumptions, the two matrices $\mathbf{L}$ and $\mathbf{S}$ can be recovered with large probability when $\lambda/\alpha = 1/\sqrt{\max\{n,p\}}$. Therefore, in our experiments, we set $\alpha = 1$ and $\lambda = 1/\sqrt{\max\{n,p\}}$.

In what concerns the fitness penalty parameter ρ, which is the single parameter that is associated with the ADMM method, the ADMM technique is known for its robustness to poor selection of its parameter. Specifically, the convergence of ADMM is guaranteed, under broad assumptions, for all positive values of its parameter [88]. Here, we set $\rho = 1$. In addition, herein paper, $\beta = 0.1\alpha$.

3.2.4 Data Description

The abundance levels of the OTUs were generated from filtered 16S rRNA gene sequencing using the naive Bayesian classifier employed by the Ribosomal Database Project (RDP) [71]. Reads with confidence below 80% were rebinned uncertain. The per-sample normalized bacterial abundance profiles were organized in a matrix called the taxonomic relative abundances matrix. RegLRSD algorithm assumes this matrix as input. Due to the unsupervised nature of RegLRSD, the labels of the samples are not required.

3.2.4.1 Dogs with Idiopathic Inflammatory Bowel Disease (IBD) Dataset

This dataset compares the fecal microbiota between 10 healthy dogs and 12 dogs diagnosed with IBD. The extracted DNA from fecal samples was sequenced by 454-pyrosequencing. OTUs were assigned based on at least 97% sequence similarity against the Greengenes reference database [89] using Quantitative Insights Into Microbial Ecology (QIIME) [90]. The sequencing data were deposited into the National Center for Biotechnology Information (NCBI)-Sequence Read Archive (SRA) under the accession number

SRP040310.

3.2.4.2 Dogs with Exocrine Pancreatic Insufficiency (EPI) Dataset

Three-days pooled fecal samples were collected from 18 healthy dogs and 7 dogs with EPI. Extracted DNA was sequenced by Illumina sequencer, and the generated sequences were analyzed using QIIME to obtain the final OTU table with at least 97% sequence similarity against the Greengenes reference database. The sequences are available in NCBI-SRA database under the accession number SRP091334.

3.2.4.3 Mouse Model of Ulcerative Colitis (UC) Dataset

This dataset represents the fecal microbiota of mice model with ulcerative colitis and control mice. The description of the samples collection, processing, and DNA extraction is described in [91]. In particular, the microbiota of 20 T-bet$^{-/-}$ x Rag2$^{-/-}$ (UC) and 10 Rag2$^{-/-}$ (control) mice was characterized using 16S data from fecal samples. The taxonomic relative abundance table is publicly available in the Supplementary Material of [42].

3.3 Results and Discussions

This section presents the comparison of the RegLRSD algorithm with the current state-of-the-art algorithms over the three metagenomic studies described in the Material and Methods Section. In particular, the proposed RegLRSD algorithm is compared with LEFSe [42] and MetaStats [43] from the statistical biomarker detection algorithms family, and MetaBoot [44] and the entropy-based filtering method from the machine learning family. Additionally, RegLRSD is compared with the standard RPCA method in order to examine the impact of adding the smoothness constraint to the original PCP problem (2.7).

3.3.1 Evaluation Criteria

The competing algorithms were evaluated using the Consistency-classification evaluation protocol shown in Fig. 2.1. The essence of this evaluation is based on generating a large number of different variations of the original dataset. Then, the evaluation metrics (i.e., classification and reproducibility) are computed by averaging the results obtained over all these variations. The details of the evaluation protocol were discussed in Section 2.3.1.

In this paper, the stability performance was visualized by presenting three types of descriptive plots. The first plot shows the average Kuncheva Index (KI) over all pairwise comparisons. The second plot provides more details about the distribution of all the KI values by presenting their histogram. An ideal algorithm in terms of stability will have the Dirac-delta distribution at KI equal to 1. This means that the algorithm generates the same set of markers over all subsamples. Practically, the more concentrated the histogram is to the right side of the plot, the more stable is the algorithm. The third plot aims to depict the stability of the ranked microbial marker lists. This is achieved by ordering all the selected markers based on their ranks. Then, a boxplot is generated for the ranks obtained in all the K subsamples for each selected marker. A perfect algorithm in the sense of stability of the ranked lists will have boxplots that are centered at the 45° line, which means that the algorithm perfectly preserves the order of the detected markers in all subsamples.

3.3.2 Simulation Setup

The classification and consistency metrics were used to measure the efficiency of the six biomarker detection algorithms in identifying potential markers. In our experiments, a random subsampling without replacement is utilized to generate 500 subsamples (i.e., $K = 500$) variations of the original dataset. Each subsample contains 80% of the samples in the original dataset (i.e., $r = 0.8$). The classification and consistency performance

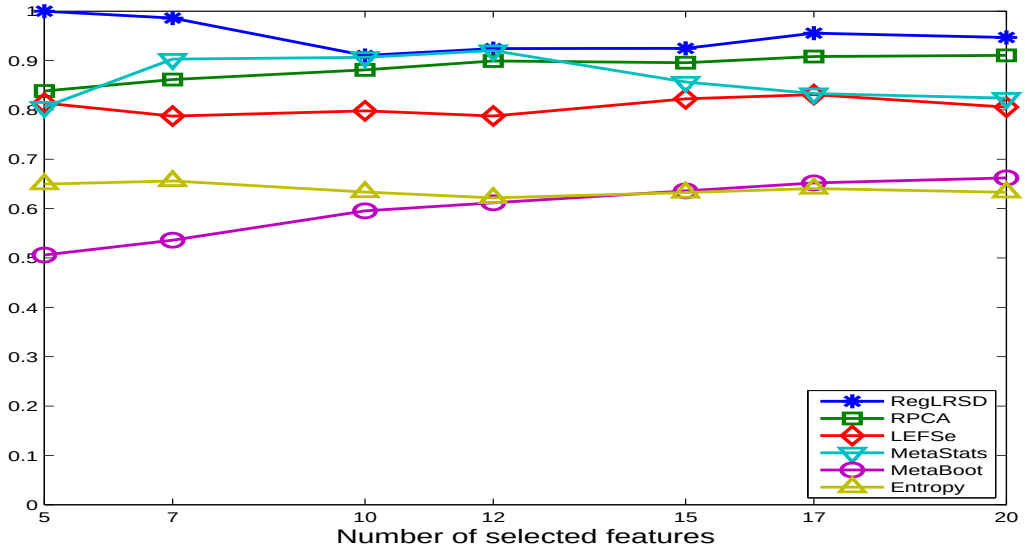

Figure 3.1: Average consistency performance measured by KI for the six biomarker discovery algorithms over the dogs from the EPI dataset.

were evaluated at different numbers of selected markers to provide further insights on the performance of the competing algorithms under varying sizes of the biomarker sets. The reported results represent the average over the 500 experiments. In our experiments, two variants of the nearest centroid classifiers were used. The first approach employed the l_1 norm as a measure of distance, while in the second approach, the l_2 norm was used. We refer to the first classifier as NCC-1 and to the second one as NCC-2.

3.3.3 Dogs with Exocrine Pancreatic Insufficiency (EPI) Dataset

The reproducibility performance in terms of the average KI stability values over all the pairwise comparisons (i.e., $K(K-1)/2 = 124750$ comparisons; $K = 500$) of the six algorithms for varying number of biomarkers from the EPI dataset is depicted in Fig. 3.1. As is clear in Fig. 3.1, RegLRSD outperforms all the other algorithms. The improvement gain of RegLRSD over the other algorithms in terms of reproducibility performance is higher at lower numbers of selected markers. This indicates that RegLRSD is more certain

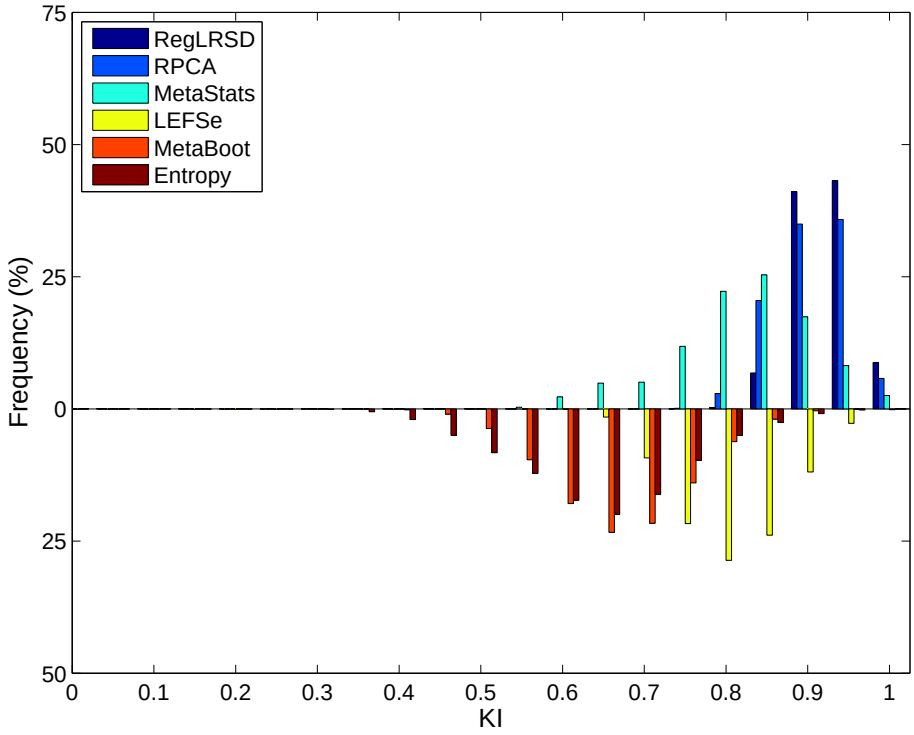

Figure 3.2: Histogram plots of all the 124750 pairwise KIs (i.e., $\frac{K(K-1)}{2} = 124750$ comparisons; $K = 500$) generated by the six biomarker discovery algorithms over the dogs from the EPI dataset.

in identifying small subsets of potential markers.

Fig. 3.2 presents the histogram of the KI index computed over the 124750 pairwise comparisons when the size of the selected biomarkers equals 20. The concentration of the histogram of RegLRSD at high KI values reveals that the RegLRSD algorithm achieves a high reproducibility performance. In particular, RegLRSD provides a stability value that is larger than or equal to 90% for almost 90% of the times. On the other hand, the other algorithms are less frequent to achieve the same stability performance. In particular, RPCA, LEFSe, and MetaStats yield a stability performance that is larger than or equal to 90% for only 75%, 15%, and 30% of the times, respectively, and less than 5% of the times for both MetaBoot and entropy-based algorithm. Moreover, the spread of the histograms

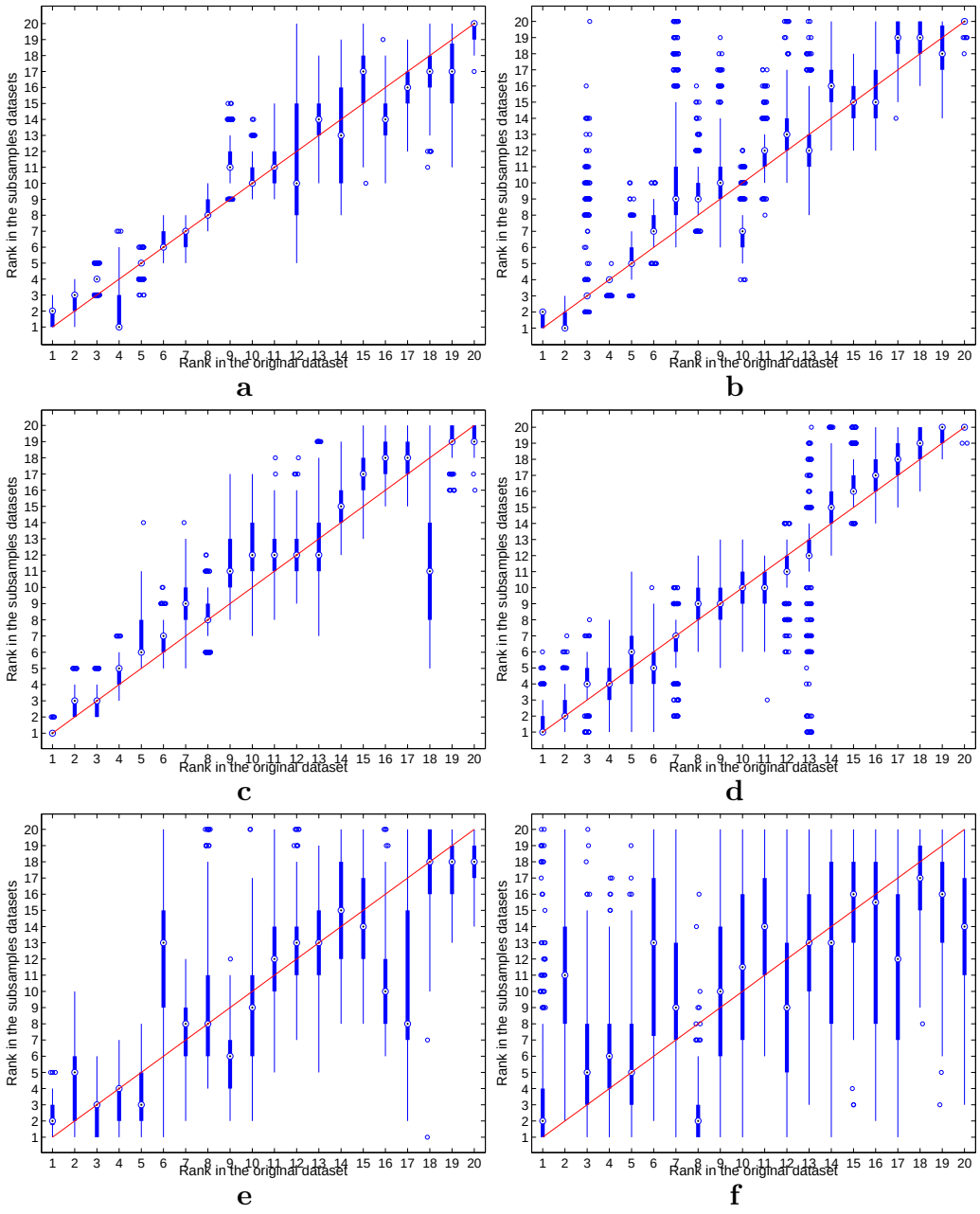

Figure 3.3: Rank boxplots in the subsamples against rank in the original data set for the six algorithms over the dogs from the EPI dataset. **(a)** RegLRSD. **(b)** RPCA. **(c)** LEFSe. **(d)** MetaStats. **(e)** MetaBoot. **(f)** Entropy.

of LEFSe, MetaStats, MetaBoot and entropy algorithms over wide range of KI values indicates a serious inconsistency problem that puts the outcomes of these algorithms under

question.

The ranking stability of the selected microbial signatures over all the $K = 500$ variations of the original dataset is depicted in Fig. 3.3. In addition to the high reproducibility performance, the RegLRSD algorithm corroborates its ability to preserve the order (i.e., rank) of the selected markers as revealed from the concentration of the boxplots of the ranks around the 45° line. The spread of the rank boxplots of the other algorithms indicates that the rank of the selected markers in these algorithms varies significantly with respect to small variations in the dataset. For example, the rank of the marker that is ranked sixth when applying the MetaBoot algorithm over the original dataset varies significantly over 500 different subsamples as cleared from Fig. 3.3.e. Specifically, the median value for all these ranks (i.e., ranks obtained in the 500 subsamples) equals 13 and the interquartile range (IQR) equals 6 (from 9 to 15). Moreover, in some subsamples, this marker was ranked first, while in other subsamples it was ranked twentieth. The classification performance of the competing algorithms is depicted in Fig. 3.4. The first column in Fig. 3.4 presents the results for the NCC-1 classifier, while the second column presents the results for the NCC-2 classifier. In general, all the algorithms provide a robust performance irrespective of the number of selected biomarkers. The identified markers by RegLRSD, LEFSe, MetaStats, and MetaBoot show high ability to distinguish between healthy and diseased samples related to EPI as revealed by the high accuracy, sensitivity and specificity of these algorithms compared to RPCA and entropy algorithms, especially when the NCC-2 is used. The better performance of RegLRSD compared to RPCA demonstrates that incorporating the prior knowledge improves the performance markedly.

Fig. 3.5 displays the top 20 identified markers by RegLRSD and their scores. RegLRSD suggests that the EPI may be characterized by the decrease in Blautia, Bacteroides, Fusobacterium, Ruminococcus genera in dogs with EPI. On the other hand, the genera, Lactobacillus, Streptococcus, Bifidobacterium exhibit a significant increase

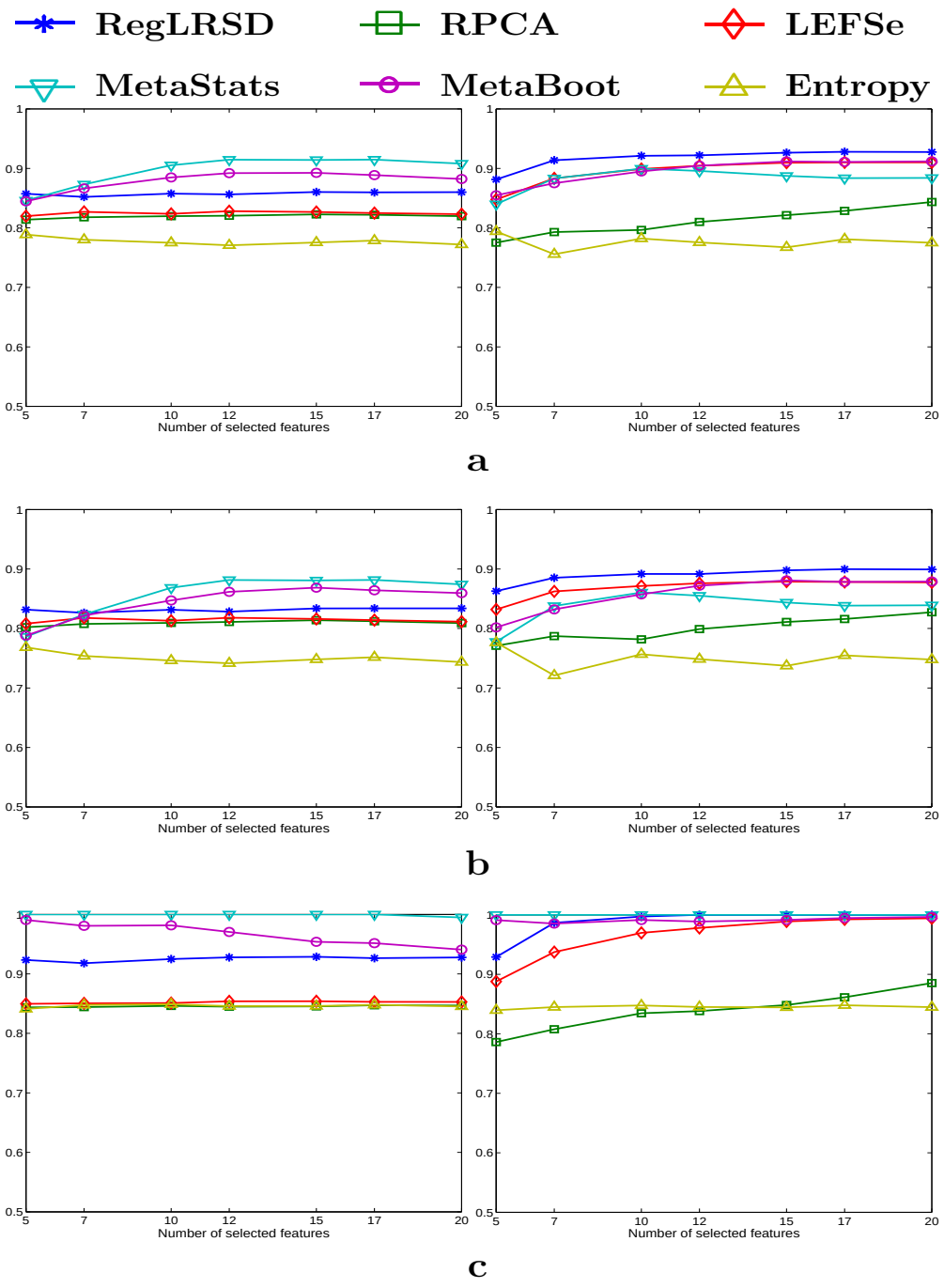

Figure 3.4: Classification performance of the six algorithms over the dogs from the EPI dataset in terms of **(a)** accuracy, **(b)** sensitivity and **(c)** specificity. The first column represents the results corresponding to the NCC-1 classifier, while the second column represents the NCC-2 classifier results.

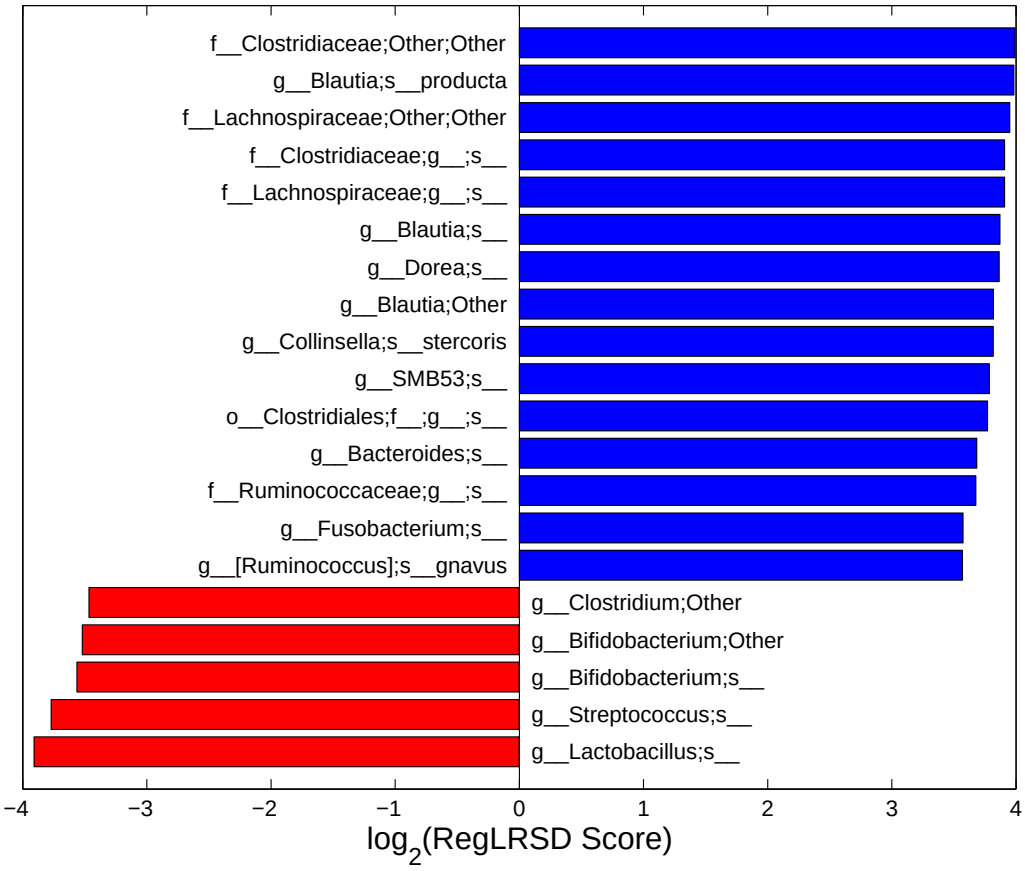

Figure 3.5: Top 20 identified biomarkers by RegLRSD in relation to the canine with EPI dataset and their RegLRSD scores. Blue: the selected bacteria exhibit an increase in their abundance level in the control samples. Red: the selected bacteria exhibit an increase in their abundance level in the EPI samples.

in their abundance levels in dogs with EPI when compared to healthy dogs. Previous studies have also showed an increase in Lactobacillus and Streptococcus abundance levels in dogs with EPI. In particular, two culture-based studies have reported an increased number of Lactobacillus and Streptococcus in the duodenum [92], jejunum and colon of dogs with EPI [93].

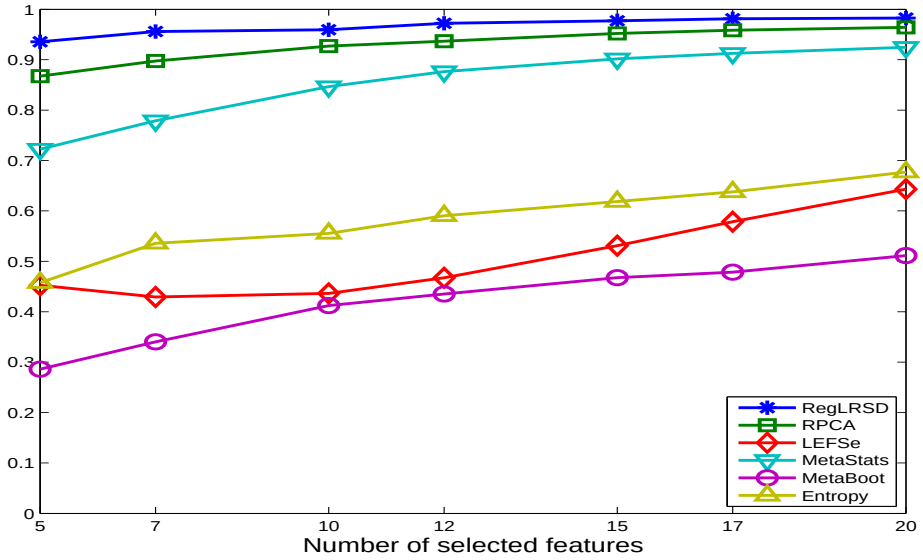

Figure 3.6: The average consistency performance measured by KI for the six biomarker discovery algorithms over the dogs from the IBD dataset.

3.3.4 Dogs with Idiopathic Inflammatory Bowel Disease (IBD) Dataset

The stability performance measured in terms of the average KI values for the six algorithms over different numbers of biomarkers is depicted in Fig. 3.6. The results in Fig. 3.6 show that RegLRSD outperforms all the other algorithms in terms of reproducibility performance. Moreover, adding the smoothing constraint in RegLRSD results in an improvement in the stability performance by almost $2-7\%$ over the standard RPCA. Noticeably, LEFSe and MetaBoot provide a poor reproducibility performance. For example, the average KI values range around $30\%-50\%$ for MetaBoot and around $40\%-65\%$ for LEFSe.

The histograms of the KI index computed over the 124750 pairwise comparisons when the size of the selected biomarkers equals 20 is depicted in Fig. 3.7. The histogram of RegLRSD illustrates the superior performance of RegLRSD as it achieves 100% stability for more than 65% of the times. RPCA and MetaStats show an adequate consistency. On the other hand, LEFSe, MetaBoot, and entropy tend to provide poor performance as their

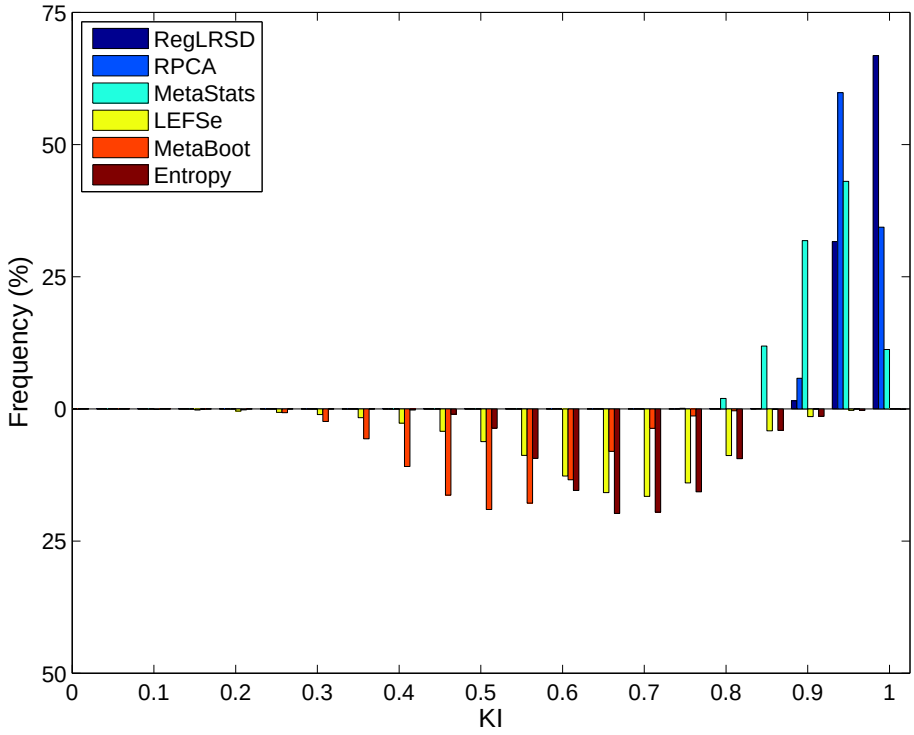

Figure 3.7: Histogram plots of all the 124750 pairwise KIs (i.e., $\frac{K(K-1)}{2} = 124750$ comparisons; $K = 500$) generated by the six biomarker discovery algorithms over the dogs from the IBD dataset.

corresponding histograms are centered at low KI values and spread over a wide range of KI values.

The ranking stability of the selected microbial signatures over all the $K = 500$ subsamples is presented in Fig. 3.8. The rank of the selected markers by RegLRSD, RPCA, and MetaBoot is more consistent against the variation in the dataset. This contrasts the performance of the LEFSe, MetaStats, and entropy-based algorithms, in which the importance (i.e., rank) of the selected features varies drastically due to adding/removing a small number of samples from the original dataset. In terms of classification performance, the RegLRSD algorithm outperforms the other algorithms especially when the NCC-2 classifier is used as revealed from Fig. 3.9. Noticeably, RegLRSD yields a significant im-

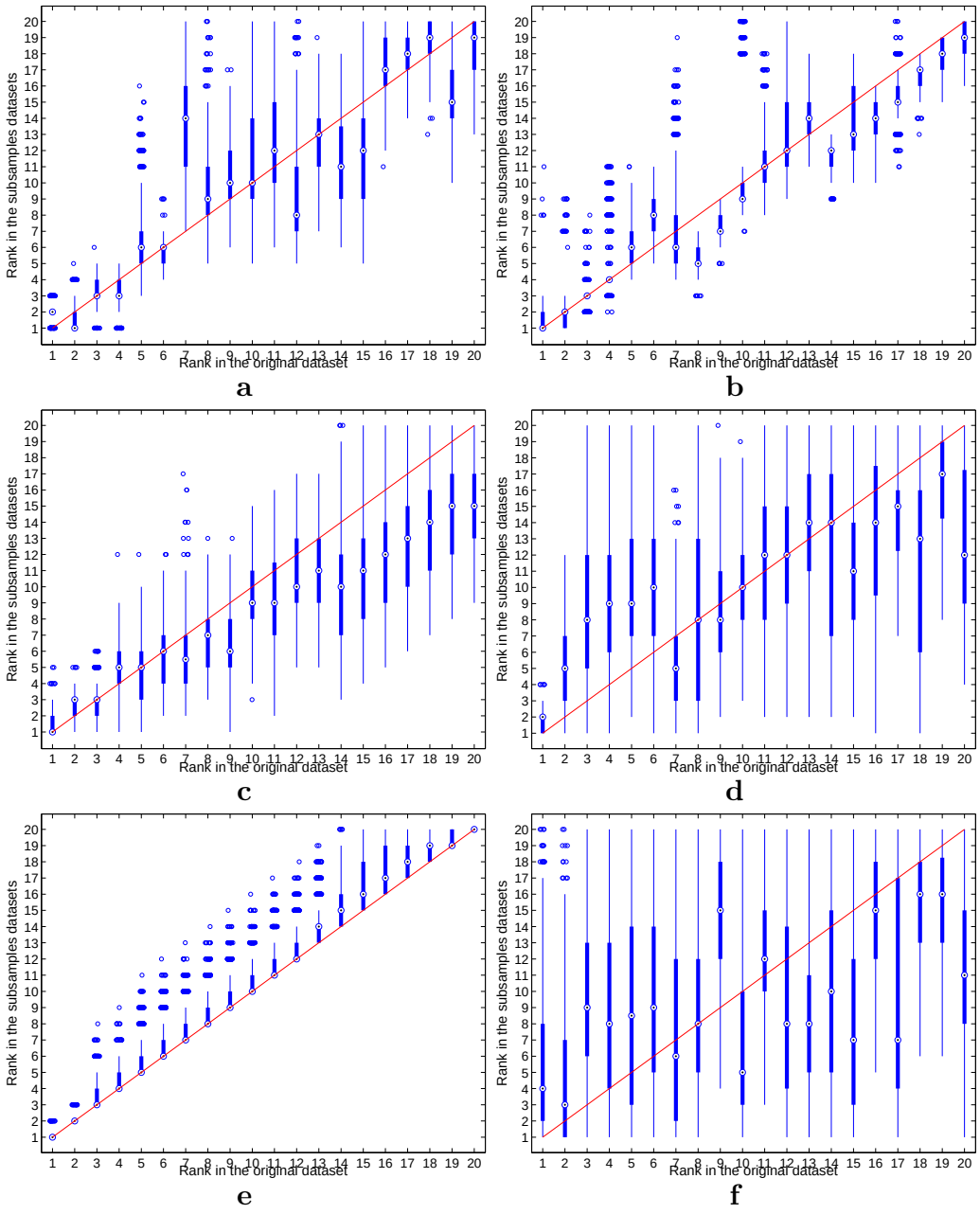

Figure 3.8: Rank boxplots in the subsamples against rank in the original data set for the six algorithms over the dogs from the IBD dataset. **(a)** RegLRSD. **(b)** RPCA. **(c)** LEFSe. **(d)** MetaStats. **(e)** MetaBoot. **(f)** Entropy.

provement over the RPCA algorithm. This reflects the efficiency of incorporating the prior knowledge information in generating more accurate results.

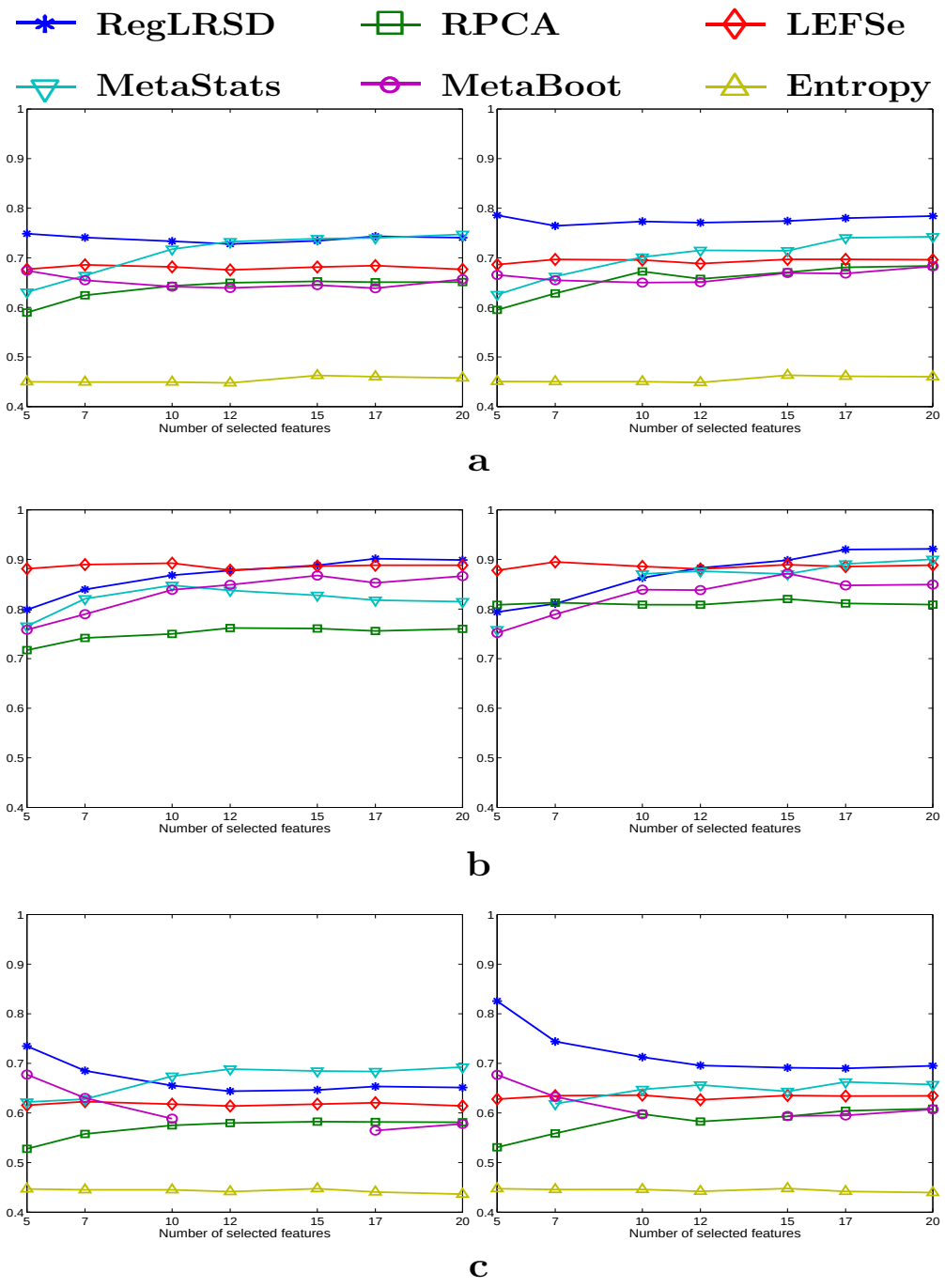

Figure 3.9: Classification performance of the six algorithms over the dogs from the IBD dataset in terms of **(a)** accuracy, **(b)** sensitivity and **(c)** specificity. The first column represents the results corresponding to the NCC-1 classifier, while the second column represents the NCC-2 classifier results.

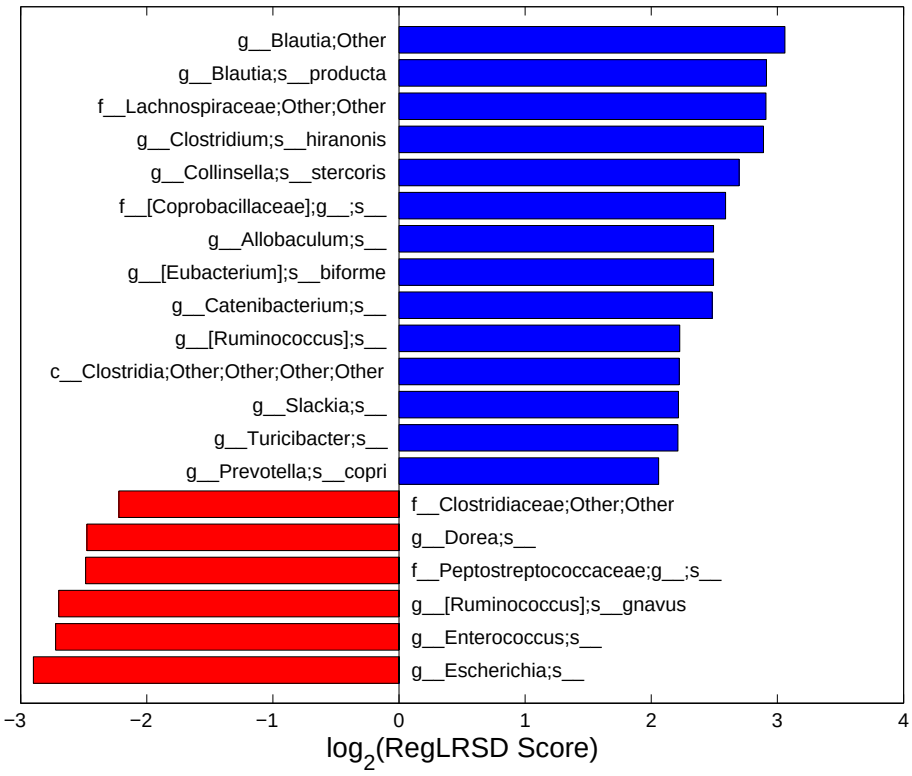

Figure 3.10: Top 20 identified biomarkers by RegLRSD in relation to the dogs with IBD dataset and their RegLRSD scores. Blue: the selected bacteria exhibit an increase in their abundance level in the control samples. Red: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

RegLRSD suggested several bacterial groups as potential markers for IBD. The top 20 detected biomarkers by the RegLRSD algorithm and their scores are displayed in Fig. 3.10. At higher phylogenetic levels, the majority of these bacterial groups belong to Firmicutes, Bacteroidetes, and Proteobacteria. In particular, the Enterobacteriaceae is the main driver for increasing the abundance level of Gammaproteobacteria in dogs with IBD. The quantitative PCR (qPCR) assays suggest that this increase is mainly due to Escherichia coli (i.e., E. coli) [94]. Several studies in human patients with IBD [95, 96] reported that E. Coli exhibits virulent potential such as adhesive capacity, invasive capacity, toxin production, and inflammatory cytokine stimulation. Similarly, the results in [97]

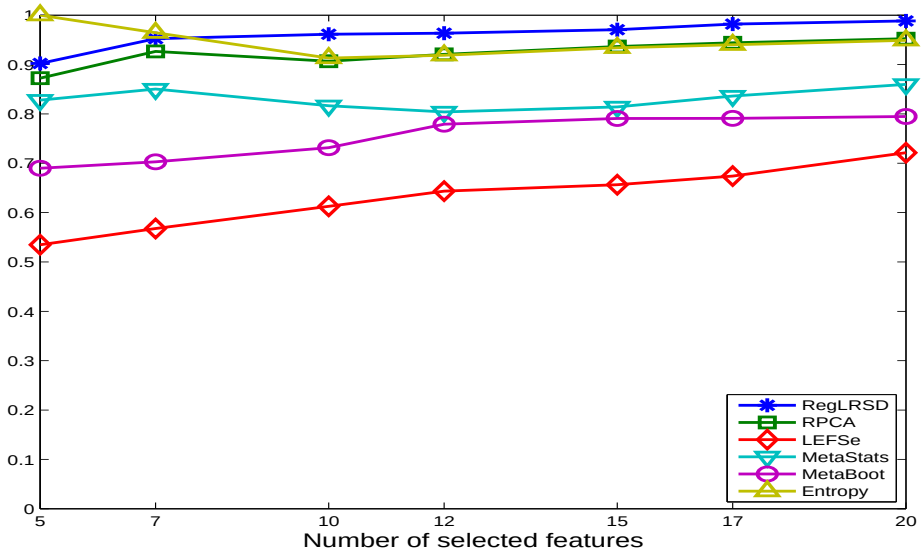

Figure 3.11: The average consistency performance measured by KI of the six biomarker discovery algorithms over the mouse model of UC dataset.

associated several adherent and invasive strains of E. Coli with granulomatous colitis in boxer dogs. RegLRSD have suggested several genera belonging to Firmicutes to be as potential markers for IBD. In particular, Blautia, Turicibacter, and Faecalibacterium were decreased in IBD. Most of these bacterial groups belong to Clostridium clusters IV and XIVa and are recognized as the major producer of several metabolites including short-chain fatty acids (SCFA). Consequently, decreasing the abundance level of these bacterial groups may impact the host health. These findings comply with previous studies in duodenal mucosal/luminal content and feces in dogs with IBD [98, 99, 75].

3.3.5 Mouse Model of Ulcerative Colitis (UC) Dataset

The average KI values over all the pairwise comparisons and their histograms when the number of selected biomarkers equals 20 are depicted in Fig. 3.11 and Fig. 3.12, respectively. Fig. 3.11 demonstrates that RegLRSD outperforms all the other algorithms and exhibits a high reproducibility performance. In particular, the improvement gain is

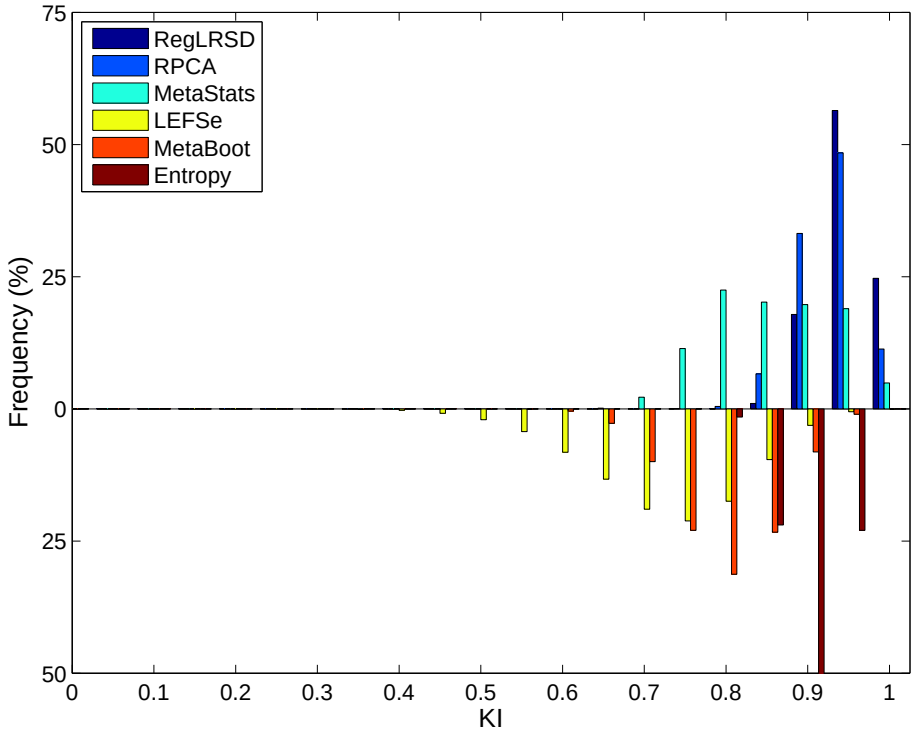

Figure 3.12: Histogram plots of all the 124750 pairwise KIs (i.e., $\frac{K(K-1)}{2} = 124750$ comparisons; $K = 500$) generated by the six biomarker discovery algorithms over the mouse model of the UC dataset.

about 5% over RPCA and entropy-based algorithm, 15% over MetaStats, $20 - 25\%$ over MetaBoot, and more than 30% over LEFSe.

The ranking stability of the selected microbial signatures over all the $K = 500$ subsamples is presented in Fig. 3.13. The results in Fig. 3.13 point a serious inconsistency problem in the performance of LEFSe, MetaStats and entropy-based algorithm. The two matrix decomposition-based algorithms (i.e., RegLRSD and RPCA) provide a comparable performance in terms of retaining the rank of the selected markers over different subsamples of the dataset.

The classification performance of the six algorithms for varying number of biomarkers from the ulcerative colitis mice model dataset is presented in Fig. 3.14. The results in

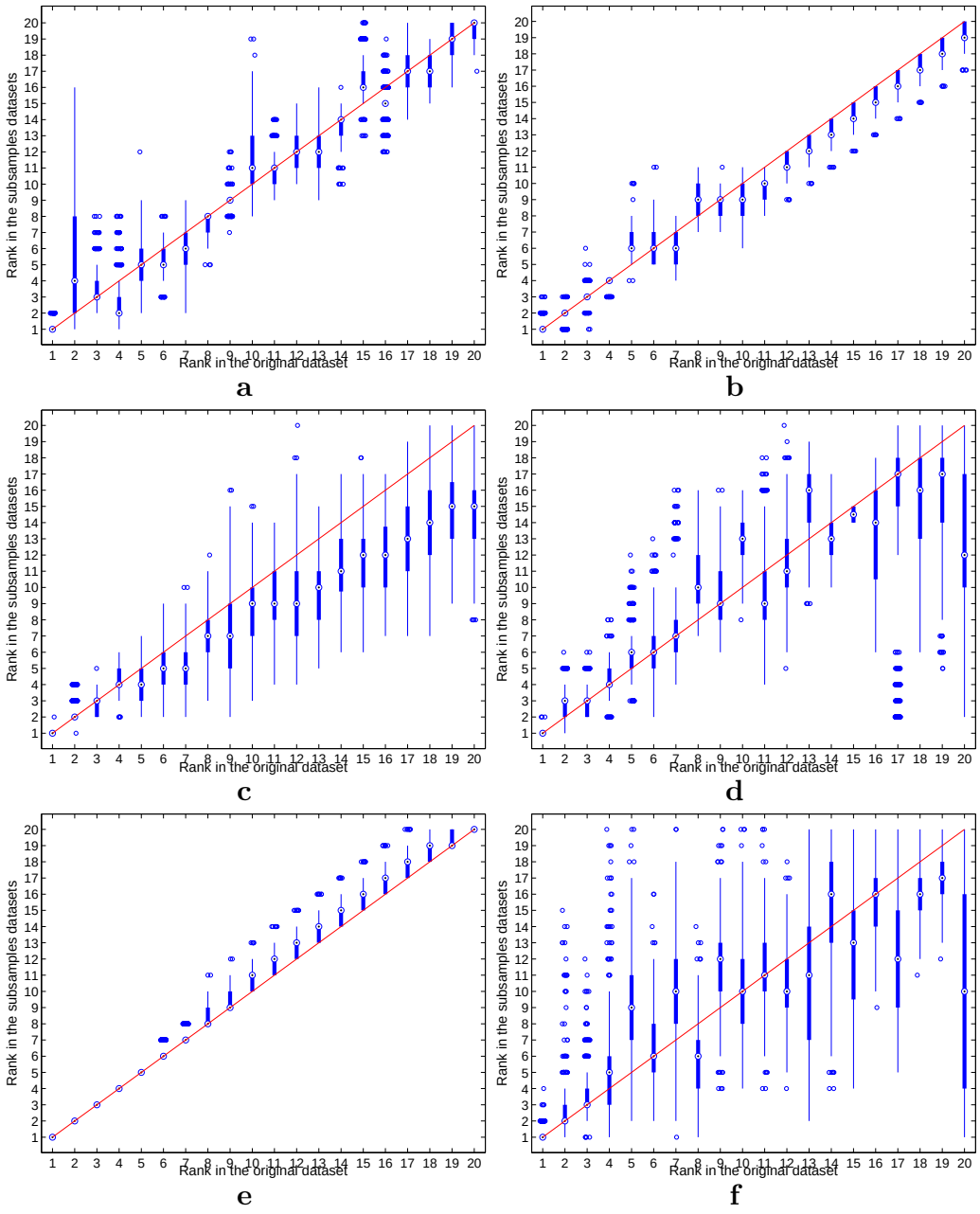

Figure 3.13: Rank boxplots in the subsamples against the rank in the original data set for the six algorithms over the mouse model of the UC dataset. **(a)** RegLRSD. **(b)** RPCA. **(c)** LEFSe. **(d)** MetaStats. **(e)** MetaBoot. **(f)** Entropy.

Fig. 3.14 point out that all the algorithms, except entropy-based algorithm, provide almost the same classification accuracy (i.e., $80 - 84\%$).

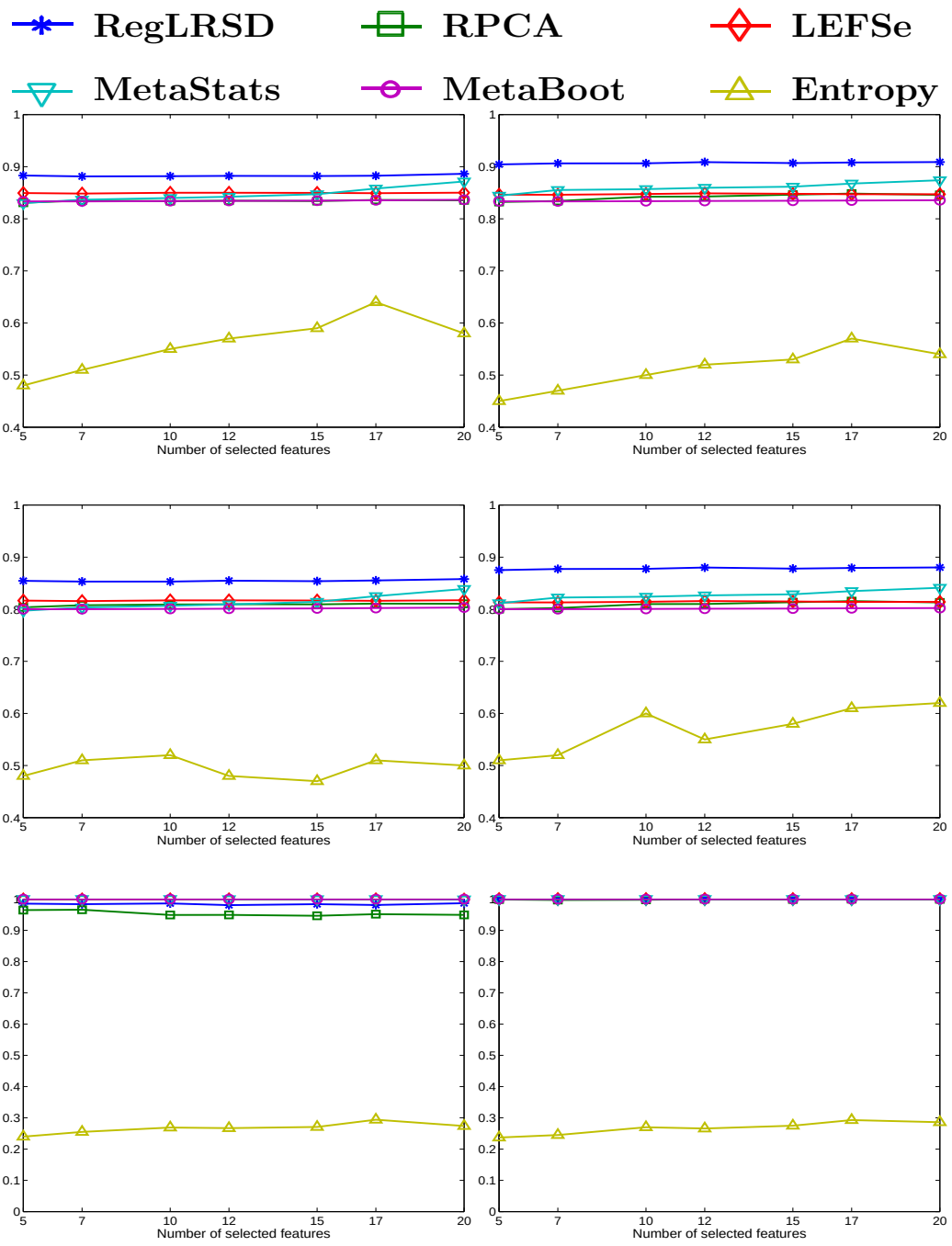

Figure 3.14: Classification performance of the six algorithms over the mouse model of the UC dataset in terms of **(a)** accuracy, **(b)** sensitivity and **(c)** specificity. The first column represents the results corresponding to the NCC-1 classifier, while the second column represents the NCC-2 classifier results.

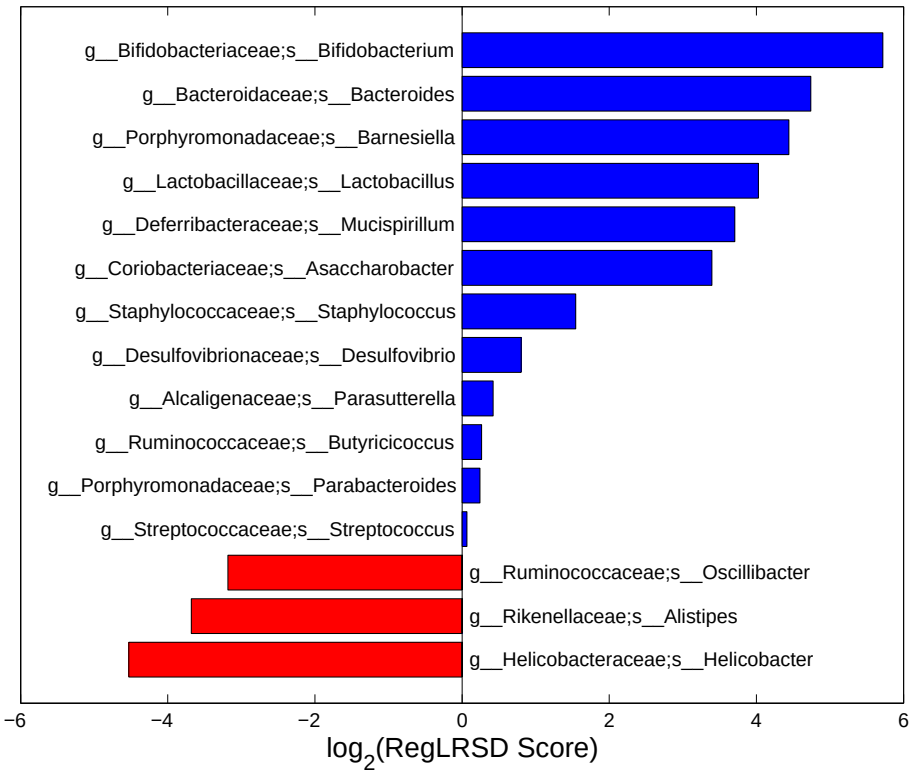

Figure 3.15: Top 15 identified biomarkers by RegLRSD in relation to the mouse model of UC dataset and their RegLRSD scores. Blue: the selected bacteria exhibit an increase in their abundance level in the control samples. Red: the selected bacteria exhibit an increase in their abundance level in the UC samples.

The top 15 identified biomarkers by the RegLRSD algorithm are listed in Fig. 3.15. The majority of these markers comply with the previous studies. For example, the authors of [81, 82] reported reduced concentrations of Lactobacillus and Bifidobacterium in colonic biopsy specimens and decreased fecal concentrations of lactobacilli and bifidobacteria in patients with active UC. The study [83] has suggested that the UC can be characterized by the decrease in the abundance levels of Bacteroides. The authors of [39] reported that the decrease in the abundance levels of acetate producer clades such as Ruminococcaceae may reduce the host ability to repair the epithelium and to regulate inflammation. This may explain the selection of Oscillibacter, which belongs to Ruminococcaceae, as a

possible marker for UC. Subjects with UC showed a significant reduction in Helicobacter pylori [79], the most widely known species of Helicobacter genus.

3.4 Summary

Recent advancements in metagenomic sequencing associated microbes with certain health and disease states of the host. Identifying potential metagenomic markers is essential for understanding biological systems and designing possible therapies for diseases. However, biomarker identification for specific diseases has been hindered by irreproducibility. This compromises practical utility in real biomedical and clinical studies that aim to identify reliable biomarkers for diagnosis, prognosis or treatment of patients. Therefore, developing robust and stable biomarker detection algorithms is crucial in order to derive solid biological conclusions and translate these findings into clinical applications.

In this paper, we developed the RegLRSD algorithm for biomarker detection. RegLRSD utilizes the convex formulation of the LRS model (2.6) to add further constraints in order to derive more accurate and consistent biological findings. Particularly, RegLRSD constrains the low-rank matrix to be smooth in order to integrate the prior knowledge that the abundance profiles of irrelevant bacteria do not exhibit a strong variation between different phenotypes in the biomarker detection process. Then we developed an efficient solution for this decomposition problem based on the alternating direction method of multipliers.

Comprehensive comparisons with state-of-the-art biomarker discovery algorithms were conducted. In particular, RegLRSD was compared with two statistical-based algorithms (i.e., LEFSe and MetaStats), two machine learning-based algorithms (MetaBoot and entropy) and a reduced form of RegLRSD in which the smoothness constraint is not considered (i.e., RPCA). The competing algorithms were tested against three realistic metagenomic datasets. The first and second datasets are in relation to healthy dogs

and dogs diagnosed with EPI and IBD, respectively. The third dataset is a mouse model of ulcerative colitis. These algorithms were evaluated in terms of classification accuracy and reproducibility performance. The simulation results show that the detected markers by RegLRSD enable discriminating metagenomic samples belonging to different phenotypes with a quite high accuracy. Moreover, RegLRSD exhibits superior consistency performance when compared to other algorithms. This renders the RegLRSD algorithm as a robust and reliable tool to identify potential metagenomic markers that may characterize the difference between samples belonging to different phenotypes.

Our simulation results demonstrate that the existing methods for metagenomic biomarker discovery present poor reproducibility performance. In particular, the spread of the histograms of LEFSe, MetaStats, MetaBoot and entropy algorithm over a wide range of KI values indicates a serious inconsistency problem that puts the outcomes of these algorithms under question. Additionally, the results reveal that the two matrix decomposition-based algorithms (i.e., RegLRSD and RPCA) are successful in providing high reproducibility and classification accuracy performance compared to the conventional statistical and machine learning-based algorithms. This validates the concept of modeling the bacterial abundance data matrix as the superposition of a low-rank matrix representing the uninformative microbes and a sparse matrix containing the abundances of informative microbes. Moreover, the improvement in the performance of RegLRSD compared to RPCA demonstrates (i) the validity of our assumption that the abundance profiles of irrelevant bacteria are smooth, and (ii) incorporating prior knowledge in the design of a biomarker detection algorithm may lead to more robust results.

4. DYSBIOSIS INDEX

4.1 Introduction

Recent metagenomic studies have associated the imbalance in the gut microbiota to the chronic inflammatory enteropathies (CE) in dogs [100, 101]. The common analysis conducted in the majority of these metagenomic studies is restricted to measure the abundance variation, evaluate the within sample diversity (α-diversity) and between sample diversity (β-diversity), and identify specific microbes or pathways that may act as potential biomarkers for the biological process at hand. Several mathematical and statistical tools have been employed in these studies such as standard hypothesis testing, supervised learning and unsupervised learning including clustering, principal component analysis (PCA) and principal coordinate analysis (PCoA).

These studies reveal a regular dysbiosis of gut microbiota due to IBD. For example, the authors in [61] assess the dysbiosis pattern due to CD in a large sample that consists of 1321 subjects, and found an association between CD and increased abundance of Enterobacteriaceae, Pasteurellacaea, Veillonellaceae, and Fusobacteriaceae, and decreased abundance in Erysipelotrichales, Bacteroidales, and Clostridiales. The results in [102] show a significant reduction in Firmicutes, in particular, Clostridium leptum in Crohn's disease (CD) subjects compared to healthy ones. In [103], the results show that IBD can be characterized by the degradation of Firmicutes and Bacteroidetes phyla. The authors in [104] identify the reduction in Faecalibacterium prausnitzii and the Firmicutes/Bacteroidetes ratio as consistent changes in IBD patients, and suggest using Faecalibacterium prausnitzii as potential biomarkers for IBD. The uniform bile-acid (BA) dysmetabolism discovered in IBD patients [105] suggests using BA as a biomarker for IBD.

These research studies are indeed useful to investigate the systematic imbalance in gut microbiota due to IBD and its potential biomarkers. However, for clinical applications, the ultimate goal is to develop a diagnostic test, which integrates the detected IBD activity indicators in an easy-to-use computational framework to measure the disease activity and to measure the response to therapy. Very few studies aim to develop a diagnostic test for IBD based on microbial markers. For example, the authors in [61] propose the Microbial Dysbiosis index (MD-index), which is defined as the logarithm of the ratio between the total abundance in organisms increased in CD and the total abundance of organisms decreased in CD. The bacteria increased in CD are Enterobacteriaceae, Pasteurellaceae, Fusobacteriaceae, Neisseriaceae, Veillonellaceae, and Gemellaceae. On the other hand, the bacteria decreased in CD are Bacteroidales, Clostridiales, Erysipelotrichales, and Bibidbacteriaceae. Another diagnostic test using fecal samples applicable for both IBD and IBS was proposed in [62] and is called 'GA-map test'. GA-map test requires DNA probes for 54 bacteria at different taxonomic levels. Since MD-index and GA-map tests are sequencing-based approaches, they suffer from two major drawbacks. The first drawback is the relatively high cost associated with the sequencing process. The second drawback is the considerable time required to receive the sequencing results. Therefore, a more efficient diagnostic test is required.

To address the challenges mentioned above, we develop a PCR-based a dysbiosis index using fecal samples to identify and track the imbalance in the gut bacterial community due to chronic enteropathies. Identifying the relative abundance of CE microbial markers using PCR technique is much cheaper and faster compared to its sequencing-based counterpart.

4.2 Material and Methods

4.2.1 Data Description

Naturally passed fecal samples were analyzed from 95 healthy dogs and 106 dogs with chronic signs of gastrointestinal disease and confirmed inflammatory changes on histopathology. Fecal samples were collected at home by the owners, immediately frozen and then shipped on ice to the laboratory. The protocol for sample collection was approved by the Texas A&M University Institutional Animal Care and Use Committee (AUP $\#2012-83$). Dogs were classified as having CE due to their chronic signs of GI disease (i.e., >3 weeks duration) and histopathologic evidence of mucosal inflammation. Clinical disease activity was scored using the canine IBD activity index (CIBDAI) with a mean of 7.6. Histological changes were predominantly of lymphoplasmacytic infiltrates, with a subset of dogs also showing eosinophilic and/or neutrophilic components.

None of the dogs received antibiotics for at least 3 weeks before sample collection. Animal information (i.e., age, weight, gender, breed) was obtained from clinical records. Also, if the owner provided the information, the exact diet (trade name and manufacturer) fed at the time of sample collection was recorded in the clinical records, and the dietary macronutrients (protein, fat, and carbohydrate content) were recorded from manufacturer's provided data on the labels.

4.2.2 Identification of the PCR Panel

DNA was extracted using the PowerFecal kit (MoBio, Carlsbad, CA). Various PCR assays were initially used to measure the abundances of selected bacterial taxa which have been previously shown to be altered in dogs with CE: total bacteria, *Proteobacteria*, *Firmicutes*, *Fusobacteria*, *Bacteroidetes*, *Ruminococcaceae*, *Bifidobacterium spp.*, *Blautia spp.*, *Faecalibacterium spp.*, *Turicibacter spp.*, *Lactobacillus spp.*, *Clostridium perfringens*, *C. hiranonis*, and *Escherichia coli*. The qPCR cycling, oligonucleotide sequences of primers

Table 4.1: Oligonucleotides primers/probes used in this study

qPCR primers/probe	Sequence (5'- 3')	Target	Annealing (C°)
Forward	GAAGGCGGCCTACTGGGCAC	Faecalibacterium	60
Reverse	GTGCAGGCGAGTTGCAGCCT		
Forward	KGGGCTCAACMCMGTATTGCGT	Fusobacteria	51
Reverse	TCGCGTTAGCTTGGGCGCTG		
Forward	TCTGATGTGAAAGGCTGGGGCTTA	Blautia	56
Reverse	GGCTTAGCCACCCGACACCTA		
Forward	CCTACGGGAGGCAGCAGT	Universal Bacteria	59
Reverse	ATTACCGCGGCTGCTGG		
Forward	CAGACGGGGACAACGATTGGA	Turicibacter	63
Reverse	TACGCATCGTCGCCTTGGTA		
Forward	GTTAATACCTTTGCTCATTGA	E. coli	55
Reverse	ACCAGGGTATCTAATCCTGTT		
Forward	AGTAAGCTCCTGATACTGTCT	C. hiranonis	50
Reverse	AGGGAAAGAGGAGATTAGTCC		
Forward	TTATTTGAAAGGGGCAATTGCT	Streptococcus	50
Reverse	GTGAACTTTCCACTCTCACAC		

and probe, and respective annealing temperatures for selected bacterial groups are shown in Table 4.1. A commercial real-time PCR thermal cycler (CFX 96 Touch^{TM} Real-Time PCR Detection System; Biorad Laboratories, Hercules, CA) was used for all qPCR assays and all samples were run in duplicate fashion.

The collected samples were divided into two groups. The first group was used as training set and consists of 36 fecal samples from healthy dogs and 55 subjects characterized with CE. The second group of samples represented an independent validation set and was composed of 59 healthy dogs and 51 dogs with CE. First, an exhaustive search wrapper feature selection method was employed to identify a smaller subset with high classification power between healthy and diseased subjects. The brute force search showed that a

subset of only seven taxa (*Faecalibacterium*, *Turicibacter*, *Streptococcus*, *E. coli*, *Blautia*, *Fusobacterium*, and *C. hiranonis*) attains a specificity and sensitivity performance close to the original fourteen bacteria. Therefore, only these seven bacteria were used to construct the dysbiosis index model.

4.2.3 Dysbiosis Index Development and Validation

To overcome the between samples variability, test and validation samples were normalized by the universal abundance level. Additionally, to guarantee the compatibility with different PCR measurement standards, the proposed index is built using the CT values rather the log-range values which varies between measurement instruments.

The dysbiosis index model is built using the nearest prototype (centroid) classifier. The nearest centroid classifier (NCC) first trains the model with labeled data in order to determine the centroid of each healthy (μ_{C_H}) and diseased (μ_{C_D}) class. Secondly, NCC assigns the test sample to the class whose centroid is closest. Geometrically, the centroid of each class can be considered as a point in a space with dimensions equal to the number of variables. The NCC classifier measures the distance between the test sample and these centroids, and assign the sample class of closest centroid. Our model employs the Euclidean distance as a measure of closeness.

The degree of dysbiosis is quantified as a single numerical value, called the Dysbiosis Index (DI), that measures the closeness (in the $l_2 - norm$) of the test sample to the mean (prototype) of each class. More formally, DI is defined as the difference between the [Euclidean distance between the test sample and the healthy class centroid] and the [Euclidean distance between the test sample and the diseased class centroid]. Mathematically, the DI of a test sample z is defined as:

$$DI(\mathbf{z}, \mu_{C_H}, \mu_{C_D}) = \|\mathbf{z} - \mu_{C_H}\|_2 - \|\mathbf{z} - \mu_{C_D}\|_2, \tag{4.1}$$

Table 4.2: Average, minimum, maximum, variance, and STD for the sensitivity and specificity values obtained by repeating 5-fold cross validation for 100 times over the training set.

Threshold	Sensitivity					Specificity				
	Average	**Min**	**Max**	**Variance**	**STD**	**Average**	**Min**	**Max**	**Variance**	**STD**
-2.00	0.87	0.5	1.00	0.01	0.11	0.89	0.57	1.00	0.01	0.11
-1.75	0.85	0.42	1.00	0.01	0.12	0.89	0.57	1.00	0.01	0.11
-1.50	0.83	0.42	1.00	0.01	0.12	0.89	0.57	1.00	0.01	0.11
-1.25	0.81	0.42	1.00	0.01	0.12	0.89	0.57	1.00	0.01	0.11
-1.00	0.80	0.42	1.00	0.01	0.12	0.89	0.57	1.00	0.01	0.11
-0.75	0.78	0.42	1.00	0.01	0.12	0.89	0.57	1.00	0.01	0.10
-0.50	0.77	0.42	1.00	0.01	0.12	0.90	0.57	1.00	0.01	0.10
-0.25	0.76	0.33	1.00	0.01	0.12	0.90	0.57	1.00	0.01	0.10
0.00	0.76	0.33	1.00	0.01	0.12	0.91	0.57	1.00	0.01	0.10
0.25	0.75	0.33	1.00	0.02	0.12	0.92	0.57	1.00	0.01	0.09
0.50	0.75	0.33	1.00	0.02	0.12	0.93	0.57	1.00	0.01	0.09
0.75	0.74	0.33	1.00	0.02	0.13	0.94	0.57	1.00	0.01	0.08
1.00	0.73	0.33	1.00	0.02	0.13	0.94	0.57	1.00	0.01	0.08
1.25	0.72	0.33	1.00	0.02	0.13	0.95	0.71	1.00	0.01	0.07
1.50	0.70	0.33	1.00	0.02	0.13	0.96	0.71	1.00	0.00	0.07
1.75	0.69	0.33	1.00	0.02	0.13	0.97	0.71	1.00	0.00	0.06
2.00	0.68	0.33	1.00	0.02	0.14	0.98	0.71	1.00	0.00	0.05

where μ_{C_D} and μ_{C_H} stand for the centroid of the diseased and healthy samples in the training set, respectively. A value of zero means that the test sample lies at equal distance from the center of both classes. The higher the DI, the more deviation of the sample from normobiosis. For example, a sample with DI equal 8 is farther away from the normobiotic reference than a sample with DI equal 2, thus the first sample is more dysbiotic than the second sample.

Table 4.3: Sensitivity and specificity performance of the DI when trained by the training set and validated by the validation set.

Threshold	Sensitivity	CI (95%)	Specificity	CI (95%)
-2	0.86	0.78-0.92	0.83	0.74-0.90
-1	0.82	0.73-0.88	0.91	0.84-0.96
0	0.74	0.65-0.82	0.95	0.89-0.98
1	0.69	0.6-0.78	0.96	0.91-0.99
2	0.63	0.53-0.72	1	0.96-1.00

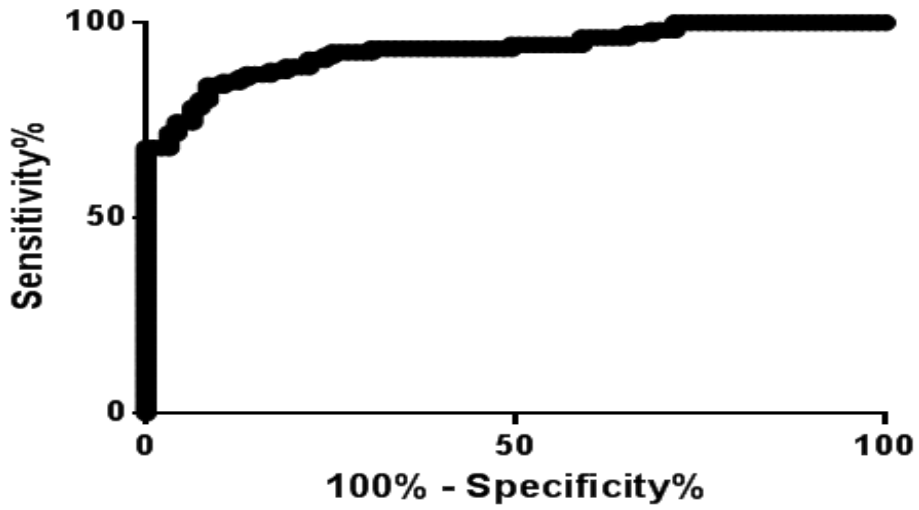

Figure 4.1: ROC curve for the DI over the independent validation set.

4.3 Results and Discussions

To evaluate the efficiency of our DI model, 5-fold stratified cross validation was conducted over the training set. To mitigate the variance in the generalization error due to small sample size, this cross-validation experiment was repeated 100 times and the average results were reported. Table 4.2 presents the basic statistics of the specificity and sensitivity values for threshold values ranging from -2 to 2. As is clear in Table 4.2, at a threshold of 0, DI achieves 76% sensitivity and 91% specificity. To assess the clinical

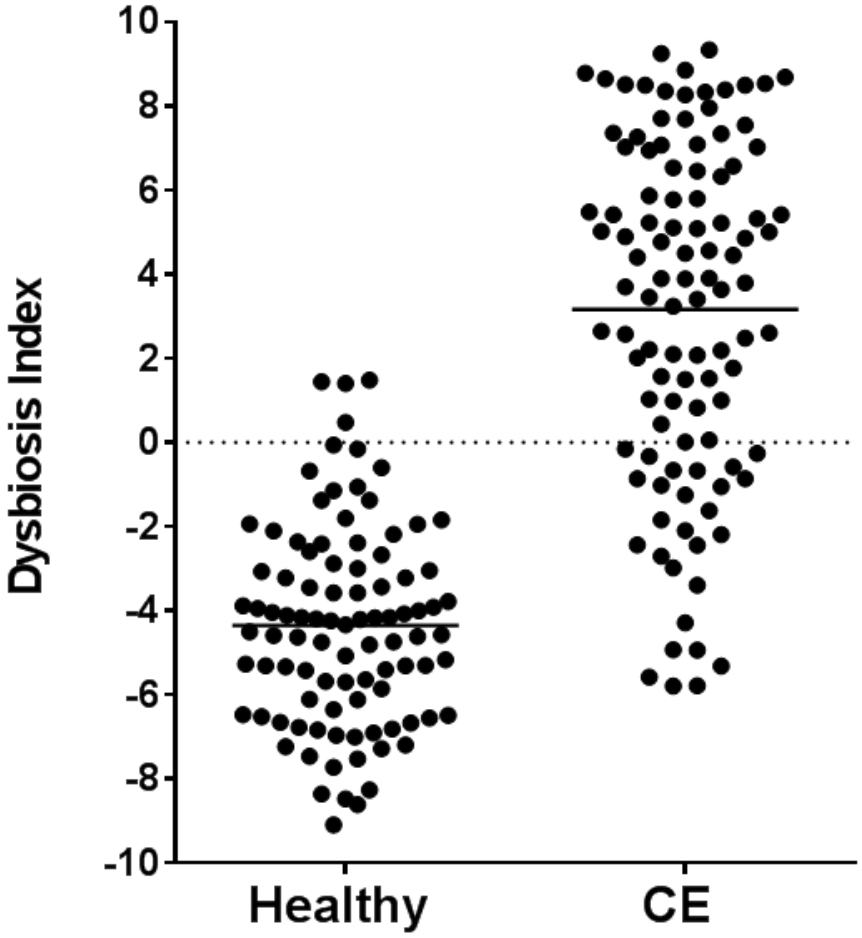

Figure 4.2: Scatter plot of the DI for all dogs in the validation set.

diagnostic performance of the DI and its capacity to track the imbalance in the bacterial population, the proposed DI was validated using an independent dataset. In particular, the training set (i.e., 36 healthy samples and 55 diseased samples) were used to train the DI model. Then, the model was tested against the validation set (i.e., 59 healthy samples and 51 diseased samples). The performance in terms of sensitivity and specificity for varying threshold values is shown in Table 4.3. For a threshold value of 0, DI yields 74% sensitivity and 95% specificity, whereas a threshold of -1 results in 82% sensitivity and 91%

specificity. The performance of the DI over the independent validation set agrees with the 5-fold cross-validation experiments conducted over the training set as is clear from Table 4.2 and Table 4.3. This demonstrates the reliability and accuracy of the DI.

To capture the performance of DI over the entire range of sensitivity/specificity values, the ROC curve is depicted in Figure 4.1. The final scatter plot of the DI for all dogs in the validation set is shown in Figure 4.2. As is clear in Figure 4.2, the DI success in characterizing health samples as the majority of healthy subjects has relatively large negative DI values. Graphs for all qPCR assays and the final DI as comparison are illustrated in Figure 4.3.

4.4 Summary

Recent metagenomic studies have associated the imbalance in the gut microbiota to the chronic enteropathies (CE) in dogs. These studies have focused on identifying the bacterial groups that may explain the systematic alteration in the intestine bacterial system due to CE. To translate these finding into clinical applications, it is required to develop a clinical diagnostic test which enables tracking the disease severity and its response to therapy. Toward this end, very few sequencing-based diagnostic tests have been proposed. In addition to the turnaround time to receive sequencing results, sequencing-based approach is costly. Therefore, this study aims to identify a PCR panel, which allows for rapid and inexpensive assessment of dysbiosis in dogs with CE. Moreover, a mathematical model is proposed to quantify these microbiota changes in a single numerical value called the Dysbiosis Index (DI).

Fecal DNA from 95 healthy dogs and 106 dogs with CE was initially analyzed for fourteen bacterial groups using quantitative PCR (qPCR) assays. These samples were grouped into two sets. The first set is a training set and it consists of 36 healthy subjects and 55 CE diseased samples. The second set is a validation set and it consists of 59

healthy subjects and 51 diseases samples. The final PCR panel consists of seven bacterial groups: Faecalibacterium, Turicibacter, E. coli, Streptococcus, Blautia, Fusobacterium, and Hiranonis. The DI was built based on the nearest centroid classifier (NCC), and it reports the degree of dysbiosis in a single numerical value that measures the closeness (in the $l_2 - norm$) of the test sample to the mean (prototype) of each class. A negative DI indicates normobiosis, whereas a positive DI indicates dysbiosis. The larger the DI, the stronger the deviation of the sample from normobiosis.

To test the DI, 100 times repetition of 5-fold cross validation experiments was conducted on the training and the average results were reported. At 0 threshold, DI achieves 76% sensitivity and 91% specificity. To validate the DI, the model was trained with the training set and tested against the validation set. For a threshold of 0, DI achieves 74% sensitivity and 95% specificity.

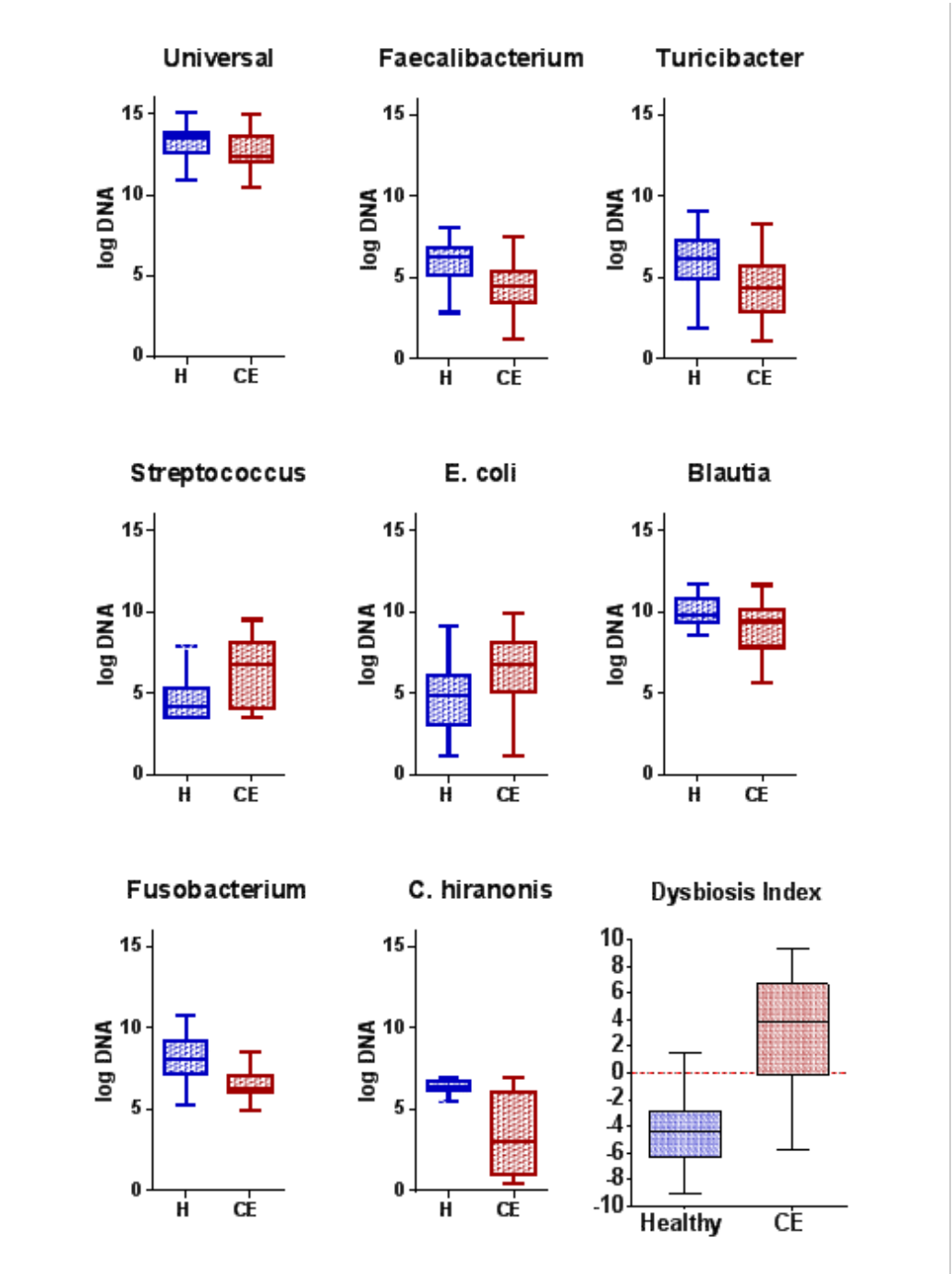

Figure 4.3: Results of the final seven qPCR assays and the final Dysbiosis Index (DI). All assays were significantly different between healthy dogs and dogs with CE ($P < 0.001$).

5. CONCLUSIONS AND FUTURE WORK

This dissertation presents a series of computational tools for the analysis of metagenomic data in order to advance our understanding of microbial communities and their relation to certain host health and disease states. These methods span two research directions: metagenomic biomarker discovery and development of dysbiosis index.

First, an evaluation protocol that accounts for the lack of knowledge of the true markers, which hinder the objective assessment of the biomarker detection algorithms, was presented. This protocol tries to mimic the the knowledge of true markers by considering the key aspects of markers. Next, a new framework for tackling the metagenomic biomarker detection problem was presented. This framework based on modeling the microbial abundance level as the superposition of: (i) a low rank matrix that models the profiles of the irrelevant features, and (ii) a sparse matrix that captures the abundance variation of the relative features. We refer to this model as the low rank-sparse (LRS) model. Then, the RPCA technique is employed to extract the sparse matrix that encodes the differential abundance information, which in turns is used to score all features. The inconsistent performance of a metagenomic biomarker selection algorithm hampers the practical utility of the detected signatures in practical applications. In order to mitigate the problem of inconsistency, the RegLRSD algorithm was developed. RegLRSD integrates the stationary nature of irrelevant microbes profiles in an attempt to yield more accurate and biological sound results.

Comprehensive comparisons with state-of-the-art biomarker discovery algorithms were conducted. In particular, the two LRS-based algorithms proposed in this dissertation (i.e., RPCA and RegLRSD) were compared with two statistical-based algorithms (i.e., LEFSe and MetaStats), and three machine learning-based algorithms (MetaBoot, entropy

and single variable classification). The competing algorithms were tested against several realistic metagenomic datasets, and were evaluated in terms of classification accuracy and reproducibility performance.

The results presented in this work demonstrate that the two matrix decomposition-based algorithms (i.e., RegLRSD and RPCA) yielded markers that enable distinguishing samples belonging to different phenotypes with high classification accuracy. In addition to the high classification power, RPCA and RegLRSD algorithms provide superior reproducibility performance compared to the conventional statistical and machine learning-based algorithms. This improved performance validates the concept of modeling the microbial abundance profiles as the superposition of a low-rank matrix representing the irrelevant microbes and a sparse matrix containing the abundances of relevant microbes. Additionally, the results show that RegLRSD presents a noticeable improvement in the performance over the RPCA algorithm. This validates (i) our assumption that the abundance level profiles of non-informative microbes can be considered smooth, and (ii) integrating prior knowledge in the design of a biomarker detection algorithm may enhance its performance.

An additional striking finding that is pointed by the results of this work is that the current state-of-the-art metagenomic biomarker discovery algorithms present a serious irreproducibility problem. In particular, these algorithms exhibit average stability values that range from low to moderate values. Additionally, their histograms of the KI values spread over a wide range of values. This severe inconsistent performance of the exiting state-of-the-art algorithms renders their findings questionable.

Another important extension treated in this dissertation was the development of a dysbiosis index for CE in canine. This dysbiosis index aims to quantify the bacterial shift in gut microbiota due to CE as a single numerical value. This dysbiosis index provides clinicians with powerful utility to diagnose the severity of IBD and measure its response

to therapies. Our results indicate that the DI is a reliable measure of clinical signs of inflammation in dogs with CE.

An interesting scenario that remains open for future investigation is to develop reliable algorithms that enable robust detection of metagenomic biomarkers from timeseries data. This is crucial to identify the bacterial groups that respond to treatment (i.e., exhibit a significant variation over the course of treatment). Integrating other OMIC data can also provide useful information that may improve the performance of the dysbiosis index.

REFERENCES

[1] J. K. Fredrickson, J. M. Zachara, D. L. Balkwill, D. Kennedy, W. L. Shu-mei, H. M. Kostandarithes, M. J. Daly, M. F. Romine, and F. J. Brockman, "Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the hanford site, washington state," *Applied and Environmental Microbiology*, vol. 70, no. 7, pp. 4230–4241, 2004.

[2] R. N. Glud, F. Wenzhöfer, M. Middelboe, K. Oguri, R. Turnewitsch, D. E. Canfield, and H. Kitazato, "High rates of microbial carbon turnover in sediments in the deepest oceanic trench on earth," *Nature Geoscience*, vol. 6, no. 4, pp. 284–288, 2013.

[3] W. B. Whitman, D. C. Coleman, and W. J. Wiebe, "Prokaryotes: the unseen majority," *Proceedings of the National Academy of Sciences*, vol. 95, no. 12, pp. 6578–6583, 1998.

[4] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, *et al.*, "A human gut microbial gene catalogue established by metagenomic sequencing," *Nature*, vol. 464, no. 7285, pp. 59–65, 2010.

[5] D. A. Relman, "Microbiology: learning about who we are," *Nature*, vol. 486, no. 7402, pp. 194–195, 2012.

[6] A. Abbott, "Scientists bust the myth that our bodies have more bacteria than human cells," *Nature News*, 2016. doi:10.1038/nature.2016.19136.

[7] S. R. Gill, M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson, "Meta-

genomic analysis of the human distal gut microbiome," *Science*, vol. 312, no. 5778, pp. 1355–1359, 2006.

[8] K. Faust and J. Raes, "Microbial interactions: from networks to models," *Nature Reviews Microbiology*, vol. 10, no. 8, pp. 538–550, 2012.

[9] V. Bucci, C. D. Nadell, and J. B. Xavier, "The evolution of bacteriocin production in bacterial biofilms," *The American Naturalist*, vol. 178, no. 6, pp. E162–E173, 2011.

[10] N. Klitgord and D. Segre, "Environments that induce synthetic microbial ecosystems," *PLoS Computational Biology*, vol. 6, no. 11, p. e1001002, 2010.

[11] A. Khosravi and S. K. Mazmanian, "Disruption of the gut microbiome as a risk factor for microbial infections," *Current Opinion in Microbiology*, vol. 16, no. 2, pp. 221–227, 2013.

[12] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower, "Microbial co-occurrence relationships in the human microbiome," *PLoS Comput Biol*, vol. 8, no. 7, pp. e1002606–e1002606, 2012.

[13] P. D. Schloss and J. Handelsman, "Metagenomics for studying unculturable microorganisms: cutting the Gordian knot," *Genome Biology*, vol. 6, no. 8, p. 229, 2005.

[14] A. Jurkowski, A. H. Reid, and J. B. Labov, "Metagenomics: a call for bringing a new science into the classroom (while it's still new)," *CBE-Life Sciences Education*, vol. 6, no. 4, pp. 260–265, 2007.

[15] P. Hugenholtz, B. M. Goebel, and N. R. Pace, "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity," *Journal of Bacteriology*, vol. 180, no. 18, pp. 4765–4774, 1998.

[16] E. J. Stewart, "Growing unculturable bacteria," *Journal of Bacteriology*, vol. 194, no. 16, pp. 4151–4160, 2012.

[17] E. A. Dinsdale, R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, *et al.*, "Functional metagenomic profiling of nine biomes," *Nature*, vol. 452, no. 7187, pp. 629–632, 2008.

[18] J. A. Gilbert, F. Meyer, and M. J. Bailey, "The future of microbial metagenomics (or is ignorance bliss?)," *The ISME Journal*, vol. 5, no. 5, p. 777, 2011.

[19] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products," *Chemistry & Biology*, vol. 5, no. 10, pp. R245–R249, 1998.

[20] M. R. Rondon, P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, *et al.*, "Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms," *Applied and Environmental Microbiology*, vol. 66, no. 6, pp. 2541–2547, 2000.

[21] N. R. Pace, "A molecular view of microbial diversity and the biosphere," *Science*, vol. 276, no. 5313, pp. 734–740, 1997.

[22] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proceedings of the National Academy of Sciences*, vol. 74, no. 11, pp. 5088–5090, 1977.

[23] J. E. Clarridge, "Impact of 16s rrna gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases," *Clinical Microbiology Reviews*, vol. 17, no. 4, pp. 840–862, 2004.

[24] J. A. Klappenbach, P. R. Saxman, J. R. Cole, and T. M. Schmidt, "rrndb: the ribosomal rna operon copy number database," *Nucleic Acids Research*, vol. 29, no. 1, pp. 181–184, 2001.

[25] J. P. McCutcheon and N. A. Moran, "Functional convergence in reduced genomes of bacterial symbionts spanning 200 my of evolution," *Genome Biology and Evolution*, vol. 2, pp. 708–718, 2010.

[26] A. Copeland, A. Lapidus, T. G. Del Rio, M. Nolan, S. Lucas, F. Chen, H. Tice, J.-F. Cheng, D. Bruce, L. Goodwin, *et al.*, "Complete genome sequence of catenulispora acidiphila type strain (id 139908 t)," *Standards in Genomic Sciences*, vol. 1, no. 2, p. 119, 2009.

[27] M. Breitbart, P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer, "Genomic analysis of uncultured marine viral communities," *Proceedings of the National Academy of Sciences*, vol. 99, no. 22, pp. 14250–14255, 2002.

[28] S. Yooseph, G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, *et al.*, "The sorcerer ii global ocean sampling expedition: expanding the universe of protein families," *PLoS Biology*, vol. 5, no. 3, p. e16, 2007.

[29] S. J. Williamson, D. B. Rusch, S. Yooseph, A. L. Halpern, K. B. Heidelberg, J. I. Glass, C. Andrews-Pfannkoch, D. Fadrosh, C. S. Miller, G. Sutton, *et al.*, "The sorcerer ii global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples," *PloS One*, vol. 3, no. 1, p. e1456, 2008.

[30] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield, "Community

structure and metabolism through reconstruction of microbial genomes from the environment," *Nature*, vol. 428, no. 6978, pp. 37–43, 2004.

[31] V. Lazarevic, K. Whiteson, S. Huse, D. Hernandez, L. Farinelli, M. Østerås, J. Schrenzel, and P. François, "Metagenomic study of the oral microbiota by illumina high-throughput sequencing," *Journal of Microbiological Methods*, vol. 79, no. 3, pp. 266–271, 2009.

[32] P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman, "Diversity of the human intestinal microbial flora," *Science*, vol. 308, no. 5728, pp. 1635–1638, 2005.

[33] B. V. Jones, F. Sun, and J. R. Marchesi, "Comparative metagenomic analysis of plasmid encoded functions in the human gut microbiome," *BMC Genomics*, vol. 11, no. 1, p. 46, 2010.

[34] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project: exploring the microbial part of ourselves in a changing world," *Nature*, vol. 449, no. 7164, p. 804, 2007.

[35] H. J. Flint, "Obesity and the gut microbiota," *Journal of Clinical Gastroenterology*, vol. 45, pp. S128–S132, 2011.

[36] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, *et al.*, "A core gut microbiome in obese and lean twins," *Nature*, vol. 457, no. 7228, pp. 480–484, 2009.

[37] V. K. Ridaura, J. J. Faith, F. E. Rey, J. Cheng, A. E. Duncan, A. L. Kau, N. W. Griffin, V. Lombard, B. Henrissat, J. R. Bain, *et al.*, "Gut microbiota from twins discordant for obesity modulate metabolism in mice," *Science*, vol. 341, no. 6150, p. 1241214, 2013.

[38] N. Larsen, F. K. Vogensen, F. Van Den Berg, D. S. Nielsen, A. S. Andreasen, B. K. Pedersen, W. A. Al-Soud, S. J. Sorensen, L. H. Hansen, and M. Jakobsen, "Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults," *PloS One*, vol. 5, no. 2, p. e9085, 2010.

[39] X. C. Morgan, T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, *et al.*, "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment," *Genome Biology*, vol. 13, no. 9, p. R79, 2012.

[40] W. Moore and L. H. Moore, "Intestinal floras of populations that have a high risk of colon cancer.," *Applied and Environmental Microbiology*, vol. 61, no. 9, pp. 3202–3207, 1995.

[41] J. Ahn, R. Sinha, Z. Pei, C. Dominianni, J. Wu, J. Shi, J. J. Goedert, R. B. Hayes, and L. Yang, "Human gut microbiome and risk of colorectal cancer," *Journal of the National Cancer Institute*, p. djt300, 2013.

[42] N. Segata, J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett, and C. Huttenhower, "Metagenomic biomarker discovery and explanation," *Genome Biology*, vol. 12, no. 6, p. R60, 2011.

[43] J. R. White, N. Nagarajan, and M. Pop, "Statistical methods for detecting differentially abundant features in clinical metagenomic samples," *PLoS Computational Biology*, vol. 5, no. 4, p. e1000352, 2009.

[44] X. Wang, X. Su, X. Cui, and K. Ning, "Metaboot: a machine learning framework of taxonomical biomarker discovery for different microbial communities based on metagenomic data," *PeerJ*, vol. 3, p. e993, 2015.

[45] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 02, pp. 185–205, 2005.

[46] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics*, vol. 19, no. 12, pp. 1484–1491, 2003.

[47] R. Simon, "Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n)," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 31–36, 2003.

[48] J. C. Wooley and Y. Ye, "Metagenomics: facts and artifacts, and computational challenges," *Journal of Computer Science and Technology*, vol. 25, no. 1, pp. 71–81, 2010.

[49] K. A. Swan, D. E. Curtis, K. B. McKusick, A. V. Voinov, F. A. Mapa, and M. R. Cancilla, "High-throughput gene mapping in caenorhabditis elegans," *Genome Research*, vol. 12, no. 7, pp. 1100–1105, 2002.

[50] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, no. 1, p. 140, 2007.

[51] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[52] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, *et al.*, "Gene-expression profiles

to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[53] A. K. Callesen, W. Vach, P. E. Jørgensen, S. Cold, O. Mogensen, T. A. Kruse, O. N. Jensen, and J. S. Madsen, "Reproducibility of mass spectrometry based protein profiles for diagnosis of breast cancer across clinical studies: a systematic review," *Journal of Proteome Research*, vol. 7, no. 4, pp. 1395–1402, 2008.

[54] B. Kleessen, A. Kroesen, H. Buhr, and M. Blaut, "Mucosal and invading bacteria in patients with inflammatory bowel disease compared with controls," *Scandinavian Journal of Gastroenterology*, vol. 37, no. 9, pp. 1034–1041, 2002.

[55] W. Strober, I. Fuss, and P. Mannon, "The fundamental basis of inflammatory bowel disease," *The Journal of Clinical Investigation*, vol. 117, no. 3, pp. 514–521, 2007.

[56] F. Guarner, "What is the role of the enteric commensal flora in ibd?," *Inflammatory Bowel Diseases*, vol. 14, pp. S83–S84, 2008.

[57] W. R. Best, J. M. Becktel, J. W. Singleton, F. Kern, *et al.*, "Development of a crohn's disease activity index," *Gastroenterology*, vol. 70, no. 3, pp. 439–444, 1976.

[58] J. S. Hyams, G. D. Ferry, F. S. Mandel, J. D. Gryboski, P. M. Kibort, B. S. Kirschner, A. M. Griffiths, A. J. Katz, R. J. Grand, J. T. Boyle, *et al.*, "Development and validation of a pediatric crohn's disease activity index.," *Journal of Pediatric Gastroenterology and Nutrition*, vol. 12, no. 4, p. 449, 1991.

[59] A. E. Jergens, C. A. Schreiner, D. E. Frank, Y. Niyo, F. E. Ahrens, P. Eckersall, T. J. Benson, and R. Evans, "A scoring index for disease activity in canine inflammatory bowel disease," *Journal of Veterinary Internal Medicine*, vol. 17, no. 3, pp. 291–297, 2003.

[60] K. Allenspach, B. Wieland, A. Gröne, and F. Gaschen, "Chronic enteropathies in dogs: evaluation of risk factors for negative outcome," *Journal of Veterinary Internal Medicine*, vol. 21, no. 4, pp. 700–708, 2007.

[61] D. Gevers, S. Kugathasan, L. A. Denson, Y. Vázquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. J. Song, M. Yassour, *et al.*, "The treatment-naive microbiome in new-onset crohn's disease," *Cell Host & Microbe*, vol. 15, no. 3, pp. 382–392, 2014.

[62] C. Casen, H. Vebø, M. Sekelja, F. Hegge, M. Karlsson, E. Ciemniejewska, S. Dzankovic, C. Frøyland, R. Nestestog, L. Engstrand, *et al.*, "Deviations in human gut microbiota: a novel diagnostic test for determining dysbiosis in patients with ibs or ibd," *Alimentary Pharmacology & Therapeutics*, vol. 42, no. 1, pp. 71–83, 2015.

[63] Z. He and W. Yu, "Stable feature selection for biomarker discovery," *Computational Biology and Chemistry*, vol. 34, no. 4, pp. 215–225, 2010.

[64] C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Küffner, and R. Zimmer, "Reliable gene signatures for microarray classification: assessment of stability and performance," *Bioinformatics*, vol. 22, no. 19, pp. 2356–2363, 2006.

[65] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[66] L. I. Kuncheva, "A stability index for feature selection.," in *Artificial Intelligence and Applications*, (Anaheim, CA, USA), pp. 421–427, ACTA Press, 2007.

[67] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[68] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[69] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in Neural Information Processing Systems*, pp. 2080–2088, 2009.

[70] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, vol. 61, no. 6, 2009.

[71] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy," *Applied and Environmental Microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007.

[72] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[73] A. Rehman, P. Lepage, A. Nolte, S. Hellmig, S. Schreiber, and S. J. Ott, "Transcriptional activity of the dominant gut mucosal microbiota in chronic inflammatory bowel disease patients," *Journal of Medical Microbiology*, vol. 59, no. 9, pp. 1114–1122, 2010.

[74] A. Swidsinski, A. Ladhoff, A. Pernthaler, S. Swidsinski, V. Loening-Baucke, M. Ortner, J. Weber, U. Hoffmann, S. Schreiber, M. Dietel, *et al.*, "Mucosal flora in inflammatory bowel disease," *Gastroenterology*, vol. 122, no. 1, pp. 44–54, 2002.

[75] G. Rossi, G. Pengo, M. Caldin, A. P. Piccionello, J. M. Steiner, N. D. Cohen, A. E. Jergens, and J. S. Suchodolski, "Comparison of microbiological, histological, and immunomodulatory parameters in response to treatment with either combination therapy with prednisone and metronidazole or probiotic vsl# 3 strains in dogs with idiopathic inflammatory bowel disease," *PloS One*, vol. 9, no. 4, p. e94699, 2014.

[76] J. Strauss, G. G. Kaplan, P. L. Beck, K. Rioux, R. Panaccione, R. DeVinney, T. Lynch, and E. Allen-Vercoe, "Invasive potential of gut mucosa-derived fusobacterium nucleatum positively correlates with ibd status of the host," *Inflammatory Bowel Diseases*, vol. 17, no. 9, pp. 1971–1978, 2011.

[77] M. Alshawaqfeh, B. Wajid, M. Guard, Y. Minamoto, J. Lidbury, J. Steiner, E. Serpedin, and J. Suchodolski, "A dysbiosis index to assess microbial changes in fecal samples of dogs with chronic enteropathy," *Journal of Veterinary Internal Medicine*, vol. 30, no. 4, p. 1536, 2016.

[78] E. de Groot, N. de Boer, M. Benninga, D. Budding, A. van Bodegraven, P. Savelkoul, and T. de Meij, "Intestinal microbiota in paediatric ulcerative colitis differs from healthy controls," *Journal of Crohn's and Colitis*, vol. 9, no. suppl 1, pp. S442–S442, 2015.

[79] X. Jin, Y. Chen, S. Chen, and Z. Xiang, "Association between helicobacter pylori infection and ulcerative colitis–a case control study from china," *International Journal of Medical Sciences*, vol. 10, no. 11, pp. 1479–1484, 2013.

[80] M. Baumgart, B. Dogan, M. Rishniw, G. Weitzman, B. Bosworth, R. Yantiss, R. H. Orsi, M. Wiedmann, P. McDonough, S. G. Kim, *et al.*, "Culture independent analysis of ileal mucosa reveals a selective increase in invasive escherichia coli of novel phylogeny relative to depletion of clostridiales in crohn's disease involving the ileum," *The ISME Journal*, vol. 1, no. 5, pp. 403–418, 2007.

[81] J. Ruseler-van Embden, W. Schouten, and L. Van Lieshout, "Pouchitis: result of microbial imbalance?," *Gut*, vol. 35, no. 5, pp. 658–664, 1994.

[82] I. Poxton, R. Brown, A. Sawyerr, and A. Ferguson, "Mucosa-associated bacterial flora of the human colon," *Journal of Medical Microbiology*, vol. 46, no. 1, pp. 85–91, 1997.

[83] M. Sasaki and J.-M. A. Klapproth, "The role of bacteria in the pathogenesis of ulcerative colitis," *Journal of Signal Transduction*, vol. 2012, 2012.

[84] T. Helleputte and P. Dupont, "Partially supervised feature selection with regularized linear models," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 409–416, ACM, 2009.

[85] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[86] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[87] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[88] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson, "Optimal parameter selection for the alternating direction method of multipliers (admm): quadratic problems," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 644–658, 2015.

[89] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, "Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb," *Applied and Environmental Microbiology*, vol. 72, no. 7, pp. 5069–5072, 2006.

[90] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, *et al.*, "Qiime allows analysis of high-throughput community sequencing data," *Nature Methods*, vol. 7, no. 5, pp. 335–336, 2010.

[91] P. Veiga, C. A. Gallini, C. Beal, M. Michaud, M. L. Delaney, A. DuBois, A. Khlebnikov, J. E. van Hylckama Vlieg, S. Punit, J. N. Glickman, *et al.*, "Bifidobacterium animalis subsp. lactis fermented milk product reduces inflammation by altering a niche for colitogenic microbes," *Proceedings of the National Academy of Sciences*, vol. 107, no. 42, pp. 18132–18137, 2010.

[92] E. Westermarck, V. Myllys, and M. Aho, "Effect of treatment on the jejunal and colonic bacterial flora of dogs with exocrine pancreatic insufficiency.," *Pancreas*, vol. 8, no. 5, pp. 559–562, 1993.

[93] K. Simpson, R. Batt, D. Jones, and D. Morton, "Effects of exocrine pancreatic insufficiency and replacement therapy on the bacterial flora of the duodenum in dogs.," *American journal of veterinary research*, vol. 51, no. 2, pp. 203–206, 1990.

[94] Y. Minamoto, C. C. Otoni, S. M. Steelman, O. Büyükleblebici, J. M. Steiner, A. E. Jergens, and J. S. Suchodolski, "Alteration of the fecal microbiota and serum metabolite profiles in dogs with idiopathic inflammatory bowel disease," *Gut microbes*, vol. 6, no. 1, pp. 33–47, 2015.

[95] R. Kotlowski, C. N. Bernstein, S. Sepehri, and D. O. Krause, "High prevalence of escherichia coli belonging to the b2+ d phylogenetic group in inflammatory bowel disease," *Gut*, vol. 56, no. 5, pp. 669–675, 2007.

[96] M. De la Fuente, L. Franchi, D. Araya, D. Díaz-Jiménez, M. Olivares, M. Álvarez-Lobos, D. Golenbock, M.-J. González, F. López-Kostner, R. Quera, *et al.*, "Escherichia coli isolates from inflammatory bowel diseases patients survive in macrophages and activate nlrp3 inflammasome," *International Journal of Medical Microbiology*, vol. 304, no. 3, pp. 384–392, 2014.

[97] K. W. Simpson, B. Dogan, M. Rishniw, R. E. Goldstein, S. Klaessig, P. L. McDonough, A. J. German, R. M. Yates, D. G. Russell, S. E. Johnson, *et al.*, "Adher-

ent and invasive escherichia coli is associated with granulomatous colitis in boxer dogs," *Infection and Immunity*, vol. 74, no. 8, pp. 4778–4792, 2006.

[98] J. S. Suchodolski, J. Camacho, and J. M. Steiner, "Analysis of bacterial diversity in the canine duodenum, jejunum, ileum, and colon by comparative 16s rrna gene analysis," *FEMS Microbiology Ecology*, vol. 66, no. 3, pp. 567–578, 2008.

[99] J. S. Suchodolski, P. G. Xenoulis, C. G. Paddock, J. M. Steiner, and A. E. Jergens, "Molecular analysis of the bacterial microbiota in duodenal biopsies from dogs with idiopathic inflammatory bowel disease," *Veterinary Microbiology*, vol. 142, no. 3, pp. 394–400, 2010.

[100] A. Jergens and K. Simpson, "Inflammatory bowel disease in veterinary medicine.," *Frontiers in Bioscience (Elite Edition)*, vol. 4, pp. 1404–1419, 2011.

[101] J. S. Suchodolski, S. E. Dowd, V. Wilke, J. M. Steiner, and A. E. Jergens, "16s rrna gene pyrosequencing reveals bacterial dysbiosis in the duodenum of dogs with idiopathic inflammatory bowel disease," *PLoS One*, vol. 7, no. 6, p. e39333, 2012.

[102] C. Manichanh, L. Rigottier-Gois, E. Bonnaud, K. Gloux, E. Pelletier, L. Frangeul, R. Nalin, C. Jarrin, P. Chardon, P. Marteau, *et al.*, "Reduced diversity of faecal microbiota in crohn's disease revealed by a metagenomic approach," *Gut*, vol. 55, no. 2, pp. 205–211, 2006.

[103] D. N. Frank, A. L. S. Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz, and N. R. Pace, "Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases," *Proceedings of the National Academy of Sciences*, vol. 104, no. 34, pp. 13780–13785, 2007.

[104] H. Sokol, P. Seksik, J. Furet, O. Firmesse, I. Nion-Larmurier, L. Beaugerie, J. Cosnes, G. Corthier, P. Marteau, and J. Doré, "Low counts of faecalibacterium

prausnitzii in colitis microbiota," *Inflammatory Bowel Diseases*, vol. 15, no. 8, pp. 1183–1189, 2009.

[105] H. Duboc, S. Rajca, D. Rainteau, D. Benarous, M.-A. Maubert, E. Quervain, G. Thomas, V. Barbu, L. Humbert, G. Despras, *et al.*, "Connecting dysbiosis, bile-acid dysmetabolism and gut inflammation in inflammatory bowel diseases," *Gut*, vol. 62, no. 4, pp. 531–539, 2013.

APPENDIX A

TOP 30 IDENTIFIED BIOMARKERS IN RELATION TO THE CANINE WITH IBD DATASET

Table A.1: Top 30 identified biomarkers by RPCA in relation to the canine IBD dataset. 1: the selected bacteria exhibit an increase in their abundance level in the control samples. 0: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

1	f__Clostridiaceae; g__Clostridium; s__
0	f__Streptococcaceae; g__Streptococcus; s__
0	f__Enterococcaceae; g__Enterococcus; s__Enterococcushaemoperoxidus
1	f__Erysipelotrichaceae; g__Catenibacterium; s__
1	f__Lachnospiraceae; g__Blautia; s__Blautiaproducta
0	f__Enterobacteriaceae; g__Serratia; s__Serratiamarcescens
0	f__Clostridiaceae; g__; s__
1	f__Lachnospiraceae; g__Ruminococcus; s__Ruminococcusgnavus
1	f__Lachnospiraceae; g__Blautia; s__
1	f__Erysipelotrichaceae; g__; s__Eubacteriumbiforme
0	f__Coriobacteriaceae; g__Collinsella; s__
1	f__Fusobacteriaceae; g__Fusobacterium; s__
0	f__Lactobacillaceae; g__Lactobacillus; s__
1	f__; g__; s__
1	f__Veillonellaceae; g__Megamonas; s__
1	f__Turicibacteraceae; g__Turicibacter; s__
1	f__Lachnospiraceae; g__; s__
1	f__Lachnospiraceae; g__Blautia; s__
0	f__; g__; s__
1	f__Lachnospiraceae; g__; s__
1	f__Erysipelotrichaceae; g__Allobaculum; s__Allobaculumstercoricanis
1	f__Ruminococcaceae; g__Faecalibacterium; s__Faecalibacteriumprausnitzii
1	f__Erysipelotrichaceae; g__Clostridium; s__
0	f__Clostridiaceae; g__Clostridium; s__
0	f__Lachnospiraceae; g__; s__
1	f__Prevotellaceae; g__Prevotella; s__
1	f__Clostridiaceae; g__Clostridium; s__
0	f__Clostridiaceae; g__; s__
1	f__Lachnospiraceae; g__; s__
1	f__Ruminococcaceae; g__; s__

Table A.2: Top 30 identified biomarkers by LEFSe in relation to the canine IBD dataset. 1: the selected bacteria exhibit an increase in their abundance level in the control samples. 0: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

0	f__Enterococcaceae; g__Enterococcus; s__Enterococcushaemoperoxidus
0	f__Coriobacteriaceae; g__Collinsella; s__
0	f__Coriobacteriaceae; g__Collinsella; s__
1	f__Erysipelotrichaceae; g__Catenibacterium; s__
1	f__Erysipelotrichaceae; g__Allobaculum; s__Allobaculumstercoricanis
1	f__Erysipelotrichaceae; g__Clostridium; s__Clostridiumcocleatum
0	f__Lactobacillaceae; g__Lactobacillus; s__Lactobacillusvaginalis
1	f__Turicibacteraceae; g__; s__
1	f__Turicibacteraceae; g__; s__
0	f__Erysipelotrichaceae; g__; s__
1	f__Erysipelotrichaceae; g__; s__
1	f__Bacteroidaceae; g__Bacteroides; s__Bacteroidescoprocola
1	f__Bacteroidaceae; g__Bacteroides; s__Bacteroidesplebeius
1	f__Clostridiaceae; g__Clostridium; s__Clostridiumhiranonis
1	f__Clostridiaceae; g__Clostridium; s__Clostridiumhiranonis
1	f__Clostridiaceae; g__Clostridium; s__Clostridiumhiranonis
0	f__Clostridiaceae; g__Clostridium; s__Clostridiumperfringens
0	f__Enterococcaceae; g__Enterococcus; s__Enterococcuscecorum
0	f__Lactobacillaceae; g__Lactobacillus; s__Lactobacillussalivarius
1	f__Fusobacteriaceae; g__J2-29; s__
1	f__Fusobacteriaceae; g__J2-29; s__
1	f__Lachnospiraceae; g__Ruminococcus; s__Ruminococcustorques
0	f__Lactobacillaceae; g__Lactobacillus; s__Lactobacillushelveticus
0	f__Bifidobacteriaceae; g__Bifidobacterium; s__Bifidobacteriumadolescentis
1	f__Erysipelotrichaceae; g__; s__Eubacteriumdolichum
0	f__Erysipelotrichaceae; g__; s__Clostridiuminnocuum
0	f__Enterobacteriaceae; g__Morganella; s__
0	f__Bifidobacteriaceae; g__Bifidobacterium; s__
0	f__Bifidobacteriaceae; g__Bifidobacterium; s__
1	f__Erysipelotrichaceae; g__Clostridium; s__Clostridiumspiroforme

Table A.3: Top 30 identified biomarkers by MetaStats in relation to the canine IBD dataset. 1: the selected bacteria exhibit an increase in their abundance level in the control samples. 0: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

0	f__Enterococcaceae; g__Melissococcus; s__Melissococcusplutonius
1	f__; g__; s__
0	f__Peptostreptococcaceae; g__Peptostreptococcus; s__Peptostreptococcusanaerobius
0	f__Erysipelotrichaceae; g__Erysipelothrix; s__Erysipelothrixrhusiopathiae
0	f__Fusobacteriaceae; g__Fusobacterium; s__
0	f__Eubacteriaceae; g__Eubacterium; s__Eubacteriumlimosum
0	f__Erysipelotrichaceae; g__Allobaculum; s__
0	f__Streptococcaceae; g__Streptococcus; s__Streptococcusminor
0	f__Bacteroidaceae; g__Bacteroides; s__Bacteroidesfragilis
0	f__Lactobacillaceae; g__Lactobacillus; s__Lactobacillussaerimneri
0	f__Erysipelotrichaceae; g__Allobaculum; s__
0	f__Bifidobacteriaceae; g__; s__
1	f__Lachnospiraceae; g__; s__
0	f__Enterobacteriaceae; g__Klebsiella; s__
0	f__Lactobacillaceae; g__Lactobacillus; s__
1	f__Lachnospiraceae; g__Epulopiscium; s__
1	f__Veillonellaceae; g__; s__
0	f__Moraxellaceae; g__Acinetobacter; s__Acinetobacterjohnsonii
1	f__Helicobacteraceae; g__; s__
0	f__Veillonellaceae; g__Veillonella; s__
0	f__Desulfovibrionaceae; g__Bilophila;s__Bilophilawadsworthia
1	f__Lachnospiraceae; g__Clostridium; s__
1	f__Lachnospiraceae; g__Coprococcus; s__
0	f__Enterococcaceae; g__Vagococcus; s__
1	f__Fusobacteriaceae; g__Fusobacterium; s__
0	f__Enterococcaceae; g__; s__
1	f__Prevotellaceae; g__Prevotella; s__
0	f__Lactobacillaceae; g__Lactobacillus; s__Lactobacillussatsumensis
1	f__Ruminococcaceae; g__; s__
1	f__; g__; s__

Table A.4: Top 30 identified biomarkers by Entropy in relation to the canine IBD dataset.
1: the selected bacteria exhibit an increase in their abundance level in the control samples.
0: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

1	f__Methanobacteriaceae; g__Methanobrevibacter; s__
1	f__Nitrososphaeraceae; g__CandidatusNitrososphaera; s__
0	f__Porphyromonadaceae; g__Porphyromonas; s__
1	f__Trebouxiophyceae; g__; s__
1	f__Bradyrhizobiaceae; g__; s__
1	f__Rhodobacteraceae; g__Paracoccus; s__
1	f__Alcaligenaceae; g__Sutterella; s__
1	f__Alcaligenaceae; g__Bordetella; s__
1	f__; g__Aquabacterium; s__
1	f__Oxalobacteraceae; g__; s__
1	f__Oxalobacteraceae; g__Herbaspirillum; s__
0	f__Pseudomonadaceae; g__Pseudomonas; s__
1	f__Enterobacteriaceae; g__; s__
0	f__Enterobacteriaceae; g__; s__
0	f__Enterobacteriaceae; g__; s__
1	f__Nannocystaceae; g__Nannocystis; s__Nannocystisexedens
1	f__Helicobacteraceae; g__Helicobacter; s__
1	f__Helicobacteraceae; g__Helicobacter; s__Helicobacterfelis
1	f__Coriobacteriaceae; g__Atopobium; s__
1	f__Geodermatophilaceae; g__Geodermatophilus; s__
1	f__Nakamurellaceae; g__; s__
1	f__Pseudonocardiaceae; g__Thermobispora; s__Thermobisporabispora
1	f__Cellulomonadaceae; g__Cellulomonas; s__
1	f__Cellulomonadaceae; g__Oerskovia; s__
1	f__Actinomycetaceae; g__Actinomyces; s__
1	f__Kineosporiaceae; g__; s__
1	f__Propionibacteriaceae; g__Propionibacterium; s__Propionibacteriumgranulosum
1	f__Nocardioidaceae; g__Nocardioides; s__
0	f__Corynebacteriaceae; g__Corynebacterium; s__
0	f__Corynebacteriaceae; g__Corynebacterium; s__

Table A.5: Top 30 identified biomarkers by binary classification in relation to the canine IBD dataset. 1: the selected bacteria exhibit an increase in their abundance level in the control samples. 0: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

0	f__Porphyromonadaceae; g__Porphyromonas; s__
0	f__Pseudomonadaceae; g__Pseudomonas; s__
0	f__Pasteurellaceae; g__; s__Pasteurellaaerogenes
0	f__Pasteurellaceae; g__; s__
0	f__Enterobacteriaceae; g__; s__
0	f__Enterobacteriaceae; g__Sodalis; s__
0	f__Enterobacteriaceae; g__; s__
0	f__Enterobacteriaceae; g__; s__
0	f__Enterobacteriaceae; g__; s__
0	f__Actinomycetaceae; g__Actinomyces; s__Actinomycesdenticolens
0	f__Corynebacteriaceae; g__Corynebacterium; s__
0	f__Corynebacteriaceae; g__Corynebacterium; s__
0	f__Nocardiaceae; g__Rhodococcus; s__Rhodococcusgloberulus
0	f__Lachnospiraceae; g__Ruminococcus; s__
0	f__Clostridiaceae; g__Clostridium; s__Clostridiumirregulare
0	f__; g__; s__
0	f__Bacillaceae; g__Anaerobacillus; s__
0	f__Bacillaceae; g__Bacillus; s__
0	f__Lactobacillaceae; g__Lactobacillus; s__Lactobacilluscrispatus
0	f__Lactobacillaceae; g__Lactobacillus; s__
0	f__Lactobacillaceae; g__Lactobacillus; s__
0	f__Carnobacteriaceae; g__Trichococcus; s__
0	f__Streptococcaceae; g__Streptococcus; s__Streptococcuspluranimalium
0	f__Clostridiaceae; g__Clostridium; s__
0	f__Clostridiaceae; g__Clostridium; s__
0	f__Clostridiaceae; g__Clostridium; s__
0	f__Helicobacteraceae; g__Helicobacter; s__
0	f__Coxiellaceae; g__Rickettsiella; s__
0	f__ClostridialesFamilyXIII.IncertaeSedis; g__; s__
0	f__Thiotrichaceae; g__Thiothrix; s__

APPENDIX B

TOP 10 IDENTIFIED BIOMARKERS IN RELATION TO THE MOUSE MODEL OF UC DATASET

Table B.1: Top 10 identified biomarkers by RPCA in relation to the mouse model of UC dataset. 1: the selected bacteria exhibit an increase in their abundance level in the control samples. 0: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

1	o__Bifidobacteriales; f__Bifidobacteriales; g__Bifidobacteriales
1	o__Bacteroidales; f__Bacteroidales; g__Bacteroidales
1	o__Lactobacillales; f__Lactobacillales; g__Lactobacillales
1	o__Bacteroidales; f__Bacteroidales; g__Bacteroidales
0	o__Bacteroidales; f__Bacteroidales; g__Bacteroidales
1	o__Deferribacterales; f__Deferribacterales; g__Deferribacterales
0	o__Campylobacterales; f__Campylobacterales; g__Campylobacterales
0	o__Clostridiales; f__Clostridiales; g__Clostridiales
0	o__Enterobacteriales; f__Enterobacteriales; g__Enterobacteriales
0	o__Lactobacillales; f__Lactobacillales; g__Lactobacillales

Table B.2: Top 10 identified biomarkers by LEFSe in relation to the mouse model of UC dataset. 1: the selected bacteria exhibit an increase in their abundance level in the control samples. 0: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

1	o__Bifidobacteriales; f__Bifidobacteriales; g__Bifidobacteriales
0	o__Campylobacterales; f__Campylobacterales; g__Campylobacterales
1	o__Bacteroidales; f__Bacteroidales; g__Bacteroidales
1	o__Deferribacterales; f__Deferribacterales; g__Deferribacterales
0	o__Bacteroidales; f__Bacteroidales; g__Bacteroidales
0	o__Clostridiales; f__Clostridiales; g__Clostridiales
0	o__Clostridiales; f__Clostridiales; g__Clostridiales
0	o__Lactobacillales; f__Lactobacillales; g__Lactobacillales
0	o__Desulfovibrionales; f__Desulfovibrionales; g__Desulfovibrionales
0	o__Enterobacteriales; f__Enterobacteriales; g__Enterobacteriales

Table B.3: Top 10 identified biomarkers by MetaStats in relation to the mouse model of UC dataset. 1: the selected bacteria exhibit an increase in their abundance level in the control samples. 0: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

0	o__Actinomycetales; f__Actinomycetales; g__Actinomycetales
0	o__Thiotrichales; f__Thiotrichales; g__Thiotrichales
1	o__Bifidobacteriales; f__Bifidobacteriales; g__Bifidobacteriales
1	o__Bifidobacteriales; f__Bifidobacteriales; g__Bifidobacteriales
0	o__Lactobacillales; f__Lactobacillales; g__Lactobacillales
1	o__Clostridiales; f__Clostridiales; g__Clostridiales
0	o__Desulfovibrionales; f__Desulfovibrionales; g__Desulfovibrionales
0	o__Coriobacteriales; f__Coriobacteriales; g__Coriobacteriales
0	o__Clostridiales; f__Clostridiales; g__Clostridiales
0	o__Clostridiales; f__Clostridiales; g__Clostridiales

Table B.4: Top 10 identified biomarkers by Entropy in relation to the mouse model of UC dataset. 1: the selected bacteria exhibit an increase in their abundance level in the control samples. 0: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

0	o__Actinomycetales; f__Actinomycetales; g__Actinomycetales
0	o__Actinomycetales; f__Actinomycetales; g__Actinomycetales
1	o__Bifidobacteriales; f__Bifidobacteriales; g__Bifidobacteriales
1	o__Bifidobacteriales; f__Bifidobacteriales; g__Bifidobacteriales
0	o__Coriobacteriales; f__Coriobacteriales; g__Coriobacteriales
0	o__Bacteroidales; f__Bacteroidales; g__Bacteroidales
0	o__Bacteroidales; f__Bacteroidales; g__Bacteroidales
0	o__Bacillales; f__Bacillales; g__Bacillales
0	o__Lactobacillales; f__Lactobacillales; g__Lactobacillales
0	o__Clostridiales; f__Clostridiales; g__Clostridiales

Table B.5: Top 10 identified biomarkers by binary classification in relation to the mouse model of UC dataset. 1: the selected bacteria exhibit an increase in their abundance level in the control samples. 0: the selected bacteria exhibit an increase in their abundance level in the IBD samples.

0	o__Thermoproteales; f__Thermoproteales; g__Thermoproteales
0	o__Methanopyrales; f__Methanopyrales; g__Methanopyrales
0	o__Acidimicrobiales; f__Acidimicrobiales; g__Acidimicrobiales
0	o__Actinomycetales; f__Actinomycetales; g__Actinomycetales
0	o__Actinomycetales; f__Actinomycetales; g__Actinomycetales
0	o__Actinomycetales; f__Actinomycetales; g__Actinomycetales
0	o__Actinomycetales; f__Actinomycetales; g__Actinomycetales
1	o__Bifidobacteriales; f__Bifidobacteriales; g__Bifidobacteriales
1	o__Bifidobacteriales; f__Bifidobacteriales; g__Bifidobacteriales
0	o__Bifidobacteriales; f__Bifidobacteriales; g__Bifidobacteriales

APPENDIX C

DETAILED DERIVATION OF RegLRSD ALGORITHM

C.1 Derivation of Remark-1

Remark-1: For any arbitrary vectors $\mathbf{u}, \mathbf{v} \in \Re^n$, and scalars $a, b \in \Re$, the following relation holds:

$$\langle a\mathbf{v} + b\mathbf{u}, \mathbf{u} \rangle = b \left\| -\frac{a}{2b}\mathbf{v} - \mathbf{u} \right\|_F^2 - \frac{a^2}{4b} \|\mathbf{v}\|_F^2 . \tag{C.1}$$

Proof:

$$\begin{aligned}
\langle a\mathbf{v} + b\mathbf{u}, \mathbf{u} \rangle &= b \left\langle \frac{a}{b}\mathbf{v} + \mathbf{u}, \mathbf{u} + \frac{a}{b}\mathbf{v} - \frac{a}{b}\mathbf{v} \right\rangle && \text{(C.2)} \\
&= b \left\langle \frac{a}{b}\mathbf{v} + \mathbf{u}, \mathbf{u} + \frac{a}{b}\mathbf{v} \right\rangle - b \left\langle \frac{a}{b}\mathbf{v} + \mathbf{u}, \frac{a}{b}\mathbf{v} \right\rangle && \text{(C.3)} \\
&= b \left[\left\langle \frac{a}{b}\mathbf{v} + \mathbf{u}, \mathbf{u} + \frac{a}{b}\mathbf{v} \right\rangle - \left\langle \frac{a}{b}\mathbf{v}, \frac{a}{b}\mathbf{v} \right\rangle - \left\langle \mathbf{u}, \frac{a}{b}\mathbf{v} \right\rangle \right] && \text{(C.4)} \\
&= b \left[\frac{a^2}{b^2}\|\mathbf{v}\|_2^2 + \frac{2a}{b}\mathbf{v}^T\mathbf{u} + \|\mathbf{u}\|_2^2 - \frac{a^2}{b^2}\|\mathbf{v}\|_2^2 - \frac{a}{b}\mathbf{v}^T\mathbf{u} \right] && \text{(C.5)} \\
&= b \left[\frac{a^2}{4b^2}\|\mathbf{v}\|_2^2 + \frac{a}{b}\mathbf{v}^T\mathbf{u} + \|\mathbf{u}\|_2^2 \right] - \frac{a^2}{4b}\|\mathbf{v}\|_2^2 && \text{(C.6)} \\
&= b \left\| \frac{a}{2b}\mathbf{v} + \mathbf{u} \right\|_2^2 - \frac{a^2}{4b}\|\mathbf{v}\|_2^2 && \text{(C.7)}
\end{aligned}$$

C.2 Derivation of the Update Step of $\mathbf{L}$

Updating $\mathbf{L}$ requires solving the following minimization problem:

$$\begin{aligned}
\mathbf{L}^{(r)} = \arg\min_{\mathbf{L}} \; & \frac{1}{2}\|\mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L}\|_F^2 + \alpha\|\mathbf{L}\|_* + \left\langle \mathbf{Z}^{(r-1)}, \mathbf{L} - \mathbf{Y}^{(r-1)} \right\rangle + \\
& \frac{\rho}{2}\|\mathbf{L} - \mathbf{Y}^{(r-1)}\|_F^2 .
\end{aligned} \tag{C.8}$$

The objective function, $\mathcal{L}_\rho(\mathbf{L}, \mathbf{Y}^{(r-1)}, \mathbf{Z}^{(r-1)})$, of the optimization problem (C.8) can be rewritten using the inner product notation as follows:

$$\begin{aligned}\mathbf{L}^{(r)} = \ \arg\min_{\mathbf{L}} \ & \frac{1}{2}\left\langle \mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L}, \mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L} \right\rangle + \alpha\|\mathbf{L}\|_* + \\ & \left\langle \mathbf{Z}^{(r-1)}, \mathbf{L} - \mathbf{Y}^{(r-1)} \right\rangle + \frac{\rho}{2}\left\langle \mathbf{L} - \mathbf{Y}^{(r-1)}, \mathbf{L} - \mathbf{Y}^{(r-1)} \right\rangle.\end{aligned} \tag{C.9}$$

The first term of $\mathcal{L}_\rho(\mathbf{L}, \mathbf{Y}^{(r-1)}, \mathbf{Z}^{(r-1)})$ can be re-expressed as:

$$\begin{aligned}\left\langle \mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L}, \mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L} \right\rangle &= \left\langle \mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L}, \mathbf{D} - \mathbf{S}^{(k)} \right\rangle - \left\langle \mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L}, \mathbf{L} \right\rangle \\ &= \left\langle \mathbf{D} - \mathbf{S}^{(k)}, \mathbf{D} - \mathbf{S}^{(k)} \right\rangle - \left\langle \mathbf{D} - \mathbf{S}^{(k)}, \mathbf{L} \right\rangle - \left\langle \mathbf{D} - \mathbf{S}^{(k)} - \mathbf{L}, \mathbf{L} \right\rangle \\ &= C_1 + \left\langle \mathbf{L} - 2(\mathbf{D} - \mathbf{S}^{(k)}), \mathbf{L} \right\rangle,\end{aligned}$$

where C_1 is a constant with respect to $\mathbf{L}$. Similarly, the third and the fourth terms of $\mathcal{L}_\rho(\mathbf{L}, \mathbf{Y}^{(r-1)}, \mathbf{Z}^{(r-1)})$ can be rewritten as:

$$\left\langle \mathbf{Z}^{(r-1)}, \mathbf{L} - \mathbf{Y}^{(r-1)} \right\rangle = \left\langle \mathbf{Z}^{(r-1)}, \mathbf{L} \right\rangle + C_2, \tag{C.10}$$

and

$$\left\langle \mathbf{L} - \mathbf{Y}^{(r-1)}, \mathbf{L} - \mathbf{Y}^{(r-1)} \right\rangle = \left\langle \mathbf{L} - 2\mathbf{Y}^{(r-1)}, \mathbf{L} \right\rangle + C_3, \tag{C.11}$$

, respectively. Variables $C_2 = \left\langle \mathbf{Z}^{(r-1)}, -\mathbf{Y}^{(r-1)} \right\rangle$ and $C_3 = \left\langle \mathbf{Y}^{(r-1)}, \mathbf{Y}^{(r-1)} \right\rangle$ are constants with respects to $\mathbf{L}$. Therefore, the optimization problem (C.8) can be rewritten as:

$$\begin{aligned}\mathbf{L}^{(r)} = \ \arg\min_{\mathbf{L}} \ & \frac{1}{2}\left\langle \mathbf{L} - 2(\mathbf{D} - \mathbf{S}^{(k)}), \mathbf{L} \right\rangle + \left\langle \mathbf{Z}^{(r-1)}, \mathbf{L} \right\rangle + \frac{\rho}{2}\left\langle \mathbf{L} - 2\mathbf{Y}^{(r-1)}, \mathbf{L} \right\rangle + \\ & \alpha\|\mathbf{L}\|_* + C_1 + C_2 + C_3.\end{aligned} \tag{C.12}$$

Rearranging the terms in (C.12) and eliminate the terms that do not depend on $\mathbf{L}$:

$$\mathbf{L}^{(r)} = \arg\min_{\mathbf{L}} \left\langle -\mathbf{D} + \mathbf{S}^{(k)} - \rho\mathbf{Y} + \mathbf{Z} + \frac{1+\rho}{2}, \mathbf{L} \right\rangle + \alpha\|\mathbf{L}\|_*. \tag{C.13}$$

Direct application of remark-1 yields the update step given by equation (3.17).