Sparse-Low Rank Matrix Decomposition Framework for Identifying Potential Biomarkers for Inflammatory Bowel Disease

Mustafa Alshawaqfeh
Department of Electrical and Computer Engineering
Texas A&M University
College Station, Texas 77843-3128
Email: mustafa.shawaqfeh@tamu.edu

Ahmad Al Kawam
Department of Electrical and Computer Engineering
Texas A&M University
College Station, Texas 77843-3128
Email: ahmad.alkawam@tamu.edu

Erchin Serpedin
Department of Electrical and Computer Engineering
Texas A&M University
College Station, Texas 77843-3128
Email: eserpedin@tamu.edu

***Abstract*—Inflammatory bowel disease (IBD) is a class of uncured chronic diseases which causes severe discomfort and in some cases could lead to life-threatening complications. Recent studies suggest a relationship between IBD and the gut microbiota. These findings reveal potential for identifying bacterial biomarkers for IBD to enable the detection and further investigation into unknown aspects of the disease. This work presents a novel method for identifying microbial biomarkers using robust principal component analysis (RPCA). Our method uses matrix decomposition to separate bacteria exhibiting a difference in abundance between healthy and diseased samples from the bacteria that have not undergone substantial change in abundance. Our method then ranks and identifies the top bacteria to be used as biomarkers. We contrast the proposed method with three well used state-of-the-art bacterial biomarker detection approaches over two datasets in relation to IBD. Our method outperforms the competing methods on the different evaluation cases.**

I. INTRODUCTION

Inflammatory bowel disease (IBD) is a class of chronic diseases in which all or part of the digestive tract becomes inflamed. IBD's symptoms usually encompass harsh weight loss, pain, fatigue, diarrhea, etc., and in critical situations, IBD could sustain life-threatening conditions. A growing body of research indicates that IBD is caused by a dysfunctional immune response to food and bacteria that are normally found in the digestive tract [1]. Due to the faulty response, the body emits white blood cells into the intestine lining, which lead to chronic sores and ulcerations.

There are two major constituents of IBD: ulcerative colitis and Crohn's disease. Ulcerative colitis (UC) represents a chronic condition which affects the intestine. In UC cases, the inner epithelial layer of the colon gives raise to tiny open sores (ulcers), which ultimately leads to abdominal unrest and perpetual colon depletion. Crohn's disease manifests through alteration of different areas of the digestive tract including the large intestine, small intestine or both spreading inflammation deep into the affected tissue. Although IBD has been studied for decades, it is still a chronic illness without a remedy and which necessitates lifetime care. These considerations call for new methods and techniques to better study the different aspects of IBD [1].

Recent studies have investigated the relationship between the IBD and the gut microbiota [2]. The gut microbiota, also known as gut flora, is the complex community of microorganisms that live in the digestive tract of many living organisms, including humans. Gut microbiota has been linked to several diseases, where the level of abundance of certain bacteria could be an indicator of the presence or absence of disease [3]. The aforementioned studies pointed out that gut bacteria are genetically associated with the sensing pathways of the intestine, which in an IBD case, could trigger the faulty immune response. Further studies link the microbial imbalance in the digestive tract to the abnormal immune response [4], [5]. These findings reveal the potential for identifying specific biomarkers for IBD among the bacterial populations. This could further our understanding of the function and development of IBD. Consequently, herein paper we propose to develop a mathematical framework to identify a set bacterial biomarkers for IBD.

In its essence, the biomarker identification problem resembles the feature selection problem in machine learning. Feature selection targets to extract a proper subset of relevant attributes from a large set of attributes. This problem has been addressed using several approaches, the most common of which are statistical, filtering, and feature transformation methods.

Statistical methods apply statistical tests for computing the p-value of each feature. The p-value represents the probability of the feature being linked to the disease. There are currently two main methods, Metastats [6] and LEFSe [7], for applying statistical assessment for microbiota biomarker discovery. For non-sparse features, Metastats employes a nonparametric t-test with permutations. It also accounts for sparse features using exact Fisher tests. On the other hand, LEFSe combines statistical insight with effect size evaluation in order to build a robust biomarker detection algorithm. The LEFS statistical analysis uses Kruskal-Wallis and Wilcoxon-Mann–Whitney testing algorithms in its process.

ISBN 978-0-9928626-7-1

In filtering approaches, the relevance of each feature is measured independently of other features. Relevance is calculated using measures such as correlation [8], [9], single classification power [10], hypothesis testing [11], [12], and several information theory measures [13], [14]. Although assessing each feature independently assumes decreased computational burden, filtering methods ignore the effects of the different combinations of features. For instance, a microbe on its own might not show strong correlation to a disease, however, it might be part of a metabolic pathway which includes other microbes whose collective effect may influence the disease.

To enable the assessment of combinations of features, several feature transformation approaches can be used. These approaches transform the original set of features into a new set such that each feature in the new set is composed of a function incorporating all the initial features. Feature transformation methods are divided into two types: supervised algorithms such as principal component analysis (PCA) and unsupervised approaches such as linear discriminant analysis-(LDA) and partial least-squares-(PLS). While the multivariate nature of transformation-based approaches accounts for complex interactions, unfortunately, it lacks the biological interpretation.

In order to utilize the multivariate advantage of transformation methods and the direct interpretation of filtering methods and statistical methods, we explore a low rank-sparse matrix decomposition framework to identify microbial biomarkers. In our proposed method we utilize the sparse matrix to detect the bacterial populations that are differentially present between diseased and control samples. In addition, the low rank matrix is used to represent the bacterial populations irrelevant to IBD. This representation relies on the fact that most of the microbes present in the sample are usually not related to the biological process studied and their levels do not change significantly between diseased and control samples. Hence, it would be reasonable to use a low-rank matrix to model their abundance level (denoted by $\mathbf{X}$). As for the few relevant microbes, their abundance levels exhibit significant variations between the diseased and control sample. This variation could be modeled using a sparse matrix (denoted by $\mathbf{S}$). Mathematically, the low rank-sparse matrix decomposition model of the bacterial abundance matrix $\mathbf{Y}$ is expressed as follows:

$$\mathbf{Y} = \mathbf{X} + \mathbf{S}. \tag{1}$$

This decomposition model concides with the decomposition problem targeted by the robust PCA (RPCA) technique. Hence, RPCA is used in decomposing $\mathbf{Y}$ as a sum involving $\mathbf{X}$ and $\mathbf{S}$. Consequently, we identify the bacterial biomarkers by recovering the matrix $\mathbf{S}$.

II. METHODS

A. Robust Principal Component Analysis

RPCA modifies the standard PCA to accounts for grossly corrupted observations. In particular, RPCA aims to decompose the data matrix into a low rank matrix $\mathbf{X}$ and a sparse matrix $\mathbf{S}$ as shown in the decomposition model (1). Recently, the authors of [15], [16] showed that $\mathbf{X}$ and $\mathbf{S}$ can be recovered *exactly*, under mild assumptions, by solving the *Principal Component Pursuit (PCP)* convex optimization problem. PCP is mathematically expressed as:

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{X}\|_* + \gamma\|\mathbf{S}\|_1 \\ &\text{subject to} \quad \mathbf{Y} = \mathbf{X} + \mathbf{S}, \end{aligned} \tag{2}$$

where $\|.\|_*$ and $\|.\|_1$ stand for the nuclear norm and l_1 norm of matrices, respectively. The nuclear norm is defined as the summation of the singular values, while the l_1 norm stands for the summation of the absolute values of the matrix elements. Variable γ represents a regularization factor that restrains the smoothness and sparseness of $\mathbf{X}$ and $\mathbf{S}$, respectively.

There are several optimization techniques in the literature for solving PCP. These methods include the iterative thresholding approach [17] and the accelerated proximal gradient algorithm [18]. This paper employs the augmented Lagrange multiplier (ALM) algorithm. The ALM algorithm solves the PCP problem through converting it into an unconstrained problem with an updated target, referred to as the *augmented Lagrangian*, which for PCP is expressed as:

$$\begin{aligned} \mathcal{L}_\beta(\mathbf{X},\mathbf{S},\mathbf{Z}) = \|\mathbf{X}\|_* + \gamma\|\mathbf{S}\|_1 + \\ \langle \mathbf{Z}, \mathbf{Y}-\mathbf{X}-\mathbf{S}\rangle + \frac{\beta}{2}\|\mathbf{Y}-\mathbf{X}-\mathbf{S}\|_F^2, \end{aligned} \tag{3}$$

where matrix $\mathbf{Z}$ captures the Lagrange multipliers. The ALM formulation presents a single regularization parameter, β, which controls the penalty of violating the equality constraints (i.e., $\mathbf{Y} = \mathbf{X} + \mathbf{S}$. Thus, the PCP problem, in the ALM framework, is formulated as

$$\begin{aligned} \text{minimize} \quad \mathcal{L}_\beta(\mathbf{X},\mathbf{S},\mathbf{Z}) = \|\mathbf{X}\|_* + \gamma\|\mathbf{S}\|_1 + \\ \langle \mathbf{Z}, \mathbf{Y}-\mathbf{X}-\mathbf{S}\rangle + \frac{\beta}{2}\|\mathbf{Y}-\mathbf{X}-\mathbf{S}\|_F^2. \end{aligned} \tag{4}$$

Solving (4) could be done iteratively. At every iteration k, two operations are performed. The first operation solves the optimization problem:

$$(\mathbf{X}_k^*, \mathbf{S}_k^*) = \arg\min_{\mathbf{X},\mathbf{S}} \quad \mathcal{L}_\beta(\mathbf{X},\mathbf{S},\mathbf{Z}_k). \tag{5}$$

The second operation updates the Lagrange multipliers through this recursion:

$$\mathbf{Z}_{k+1} = \mathbf{Z}_k + \beta\left(\mathbf{Y} - \mathbf{X}_k - \mathbf{S}_k\right). \tag{6}$$

However, a joint optimal solution for (5) cannot be efficiently found. Therefore, using alternating optimization could offer an efficient practical implementation. The essence of alternating-based optimization techniques is to decompose the optimization problem into smaller sub-problems. Each sub-problem solves only for one variable while fixing the remaining variables constants. The alternating optimization formulation for problem (5) assumes the following two sub-problems. The first sub-problem solves $\min_{\mathbf{X}} \ \mathcal{L}_\beta(\mathbf{X},\mathbf{S},\mathbf{Z}_k)$ as a function of $\mathbf{X}$, while $\mathbf{S}$ is set fixed. The second sub-problem solves $\min_{\mathbf{S}} \ \mathcal{L}_\beta(\mathbf{X},\mathbf{S},\mathbf{Z}_k)$ with respect to $\mathbf{S}$, while $\mathbf{X}$ is fixed. This

ISBN 978-0-9928626-7-1 © EURASIP 2017

approach relies on the fact that the two sub-problems admit closed form solutions. Consider $\mathcal{S}_\tau : \Re \to \Re$ represents the shrinkage operator:

$$\mathcal{S}_\tau(x) = \text{sgn}(x)\text{max}(|x| - \tau, 0), \tag{7}$$

with $\tau \geq 0$ denoting a positive threshold. The extension of the shrinkage operator is achieved by employing it to the matrix's entries. Thus,

$$\begin{aligned} \mathbf{S}^* &= \arg\min_{\mathbf{S}} \quad \mathcal{L}_\beta(\mathbf{X}, \mathbf{S}, \mathbf{Z}) \\ &= \mathcal{S}_{\gamma\beta^{-1}}(\mathbf{Y} - \mathbf{X} + \beta^{-1}\mathbf{Z}). \end{aligned} \tag{8}$$

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

$$\mathcal{D}_\tau(\mathbf{M}) = \mathbf{U}\mathcal{S}_\tau(\Sigma)\mathbf{V}^T, \tag{9}$$

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

$$\begin{aligned} \mathbf{X}^* &= \arg\min_{\mathbf{X}} \quad \mathcal{L}_\beta(\mathbf{X}, \mathbf{S}, \mathbf{Z}) \\ &= \mathcal{D}_{\beta^{-1}}(\mathbf{Y} - \mathbf{S} + \beta^{-1}\mathbf{Z}). \end{aligned} \tag{10}$$

B. RPCA framework for assessing the differentially abundant bacteria.

We divide the bacterial biomarker identification problem into two steps. In the first step we apply RPCA to separate the bacteria exhibiting differential abundance between diseased and control samples from the bacteria with non-differential abundance. The second step employs a scoring process to the differentially abundant bacteria to identify the top m bacteria to be selected as biomarkers.

Let's inspect the bacterial abundance level matrix $\mathbf{Y} \in \Re_+^{p\times n}$. The abundance levels of all p bacterial groups in the j^{th} sample are represented using the j^{th} column of $\mathbf{Y}$ (denoted by $\mathbf{y}_i$). The i^{th} row of $\mathbf{Y}$ represents the abundance degrees of the i^{th} microbe in all n tests. Since we expect the potential bacterial markers to have different abundance levels between samples from different phenotypes, we can modele these biomarkers using a sparse matrix, $\mathbf{S}$ expressed as

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

The nonzero entries of $\mathbf{S}$ can take either positive or negative values depending on whether the bacteria's abundance increased or decreased in response to IBD. Therefore, absolute values of $\mathbf{S}$ are used to identify the bacteria with differential abundance. In particular, the $i'th$ bacteria is assigned a score equals to the absolute values of the $i'th$ row in $\mathbf{S}$. Therefore, we can express the scoring vector ($\mathbf{v}$) as:

$$\mathbf{v} = \left[\sum_{j=1}^{n} |s_{1j}|, \dots, \sum_{j=1}^{n} |s_{pj}| \right]^T. \tag{12}$$

Large scores represent microbes with large variation in abundance between the two groups. The m microbes with the highest scores are selected as biomarkers.

III. EXPERIMENTAL RESULTS

A. Data description

To evaluate RPCA's ability to identify bacterial biomarkers for IBD, we tested our approach on two datasets in two different model organisms. The first dataset is composed of Canine subjects having IBD. Alternatively, the second dataset is composed of mice having UC, which is a type of IBD. For both datasets, the bacterial abundances are produced using 16S rRNA gene sequencing reads. We construct the relative abundance matrix using the per-sample normalized read counts. A detail description of the datasets is given next:

1) Canine IBD dataset: Microbiota information is extracted from the fecal samples of 89 healthy dogs and 79 dogs showing signs of chronic gastrointestinal (GI) illness. The dogs exhibiting traces of chronic GI disease were classified with idiopathic IBD using the criteria specified by the World Small Animal Veterinary Association (WSAVA). These datasets are available at the link: https://qiita.ucsd.edu/study/description/833.

2) Ulcerative colitis (UC) in mice dataset: Similar to the canine dataset, microbiota information was extracted from the fecal samples of 20 T-bet$^{-/-}$ x Rag2$^{-/-}$ mice with ulcerative colitis and 10 Rag2$^{-/-}$ control mice. This dataset is provided in [7].

B. Evaluation

Identifying high quality bacterial biomarkers enhances the ability of separating diseased from healthy subjects. Therefore, to evaluate our method's performance, we applied our method to the datasets described above to identify the top bacterial biomarkers. We then used these biomarkers to classify the canine and mouse subjects into two classes (healthy and diseased, respectively) to assess the discriminative power of the detected markers. Furthermore, we applied two different classifiers, k-nearest neighbors (kNN) [19] and nearest centroid classifier (NCC) [20] with l_1 norm as a measure of distance (NCC-L1) in order to decrease the assess the method's performance across different classification methods.

Evaluating classification performance could be typically achieved through calculating the classification accuracy. This is achieved by assessing the percent of situations correctly classified in both classes. However, the typical accuracy measure could suffer a major shortcoming in the case of imbalanced class distribution where its value becomes dominated by the accuracy value for the majority class.

To overcome this issue, we used a balanced accuracy (BA) measure, expressed as the mean of the accuracy received from each class (i.e., average of sensitivity and specificity) [21]. Mathematically, BA is expressed as:

$$\text{BA} = \frac{TP}{P} + \frac{TN}{N} = \frac{sensitivity + specificity}{2}, \tag{13}$$

where TP represents the true positives, TN stands for true negatives, and FN denote the false negatives. The equal weighting

ISBN 978-0-9928626-7-1 © EURASIP 2017

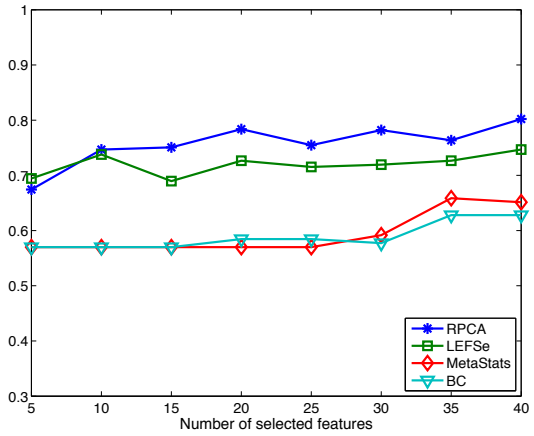

Fig. 1. Balanced accuracy of the four algorithms using kNN classification over the canine IBD dataset.

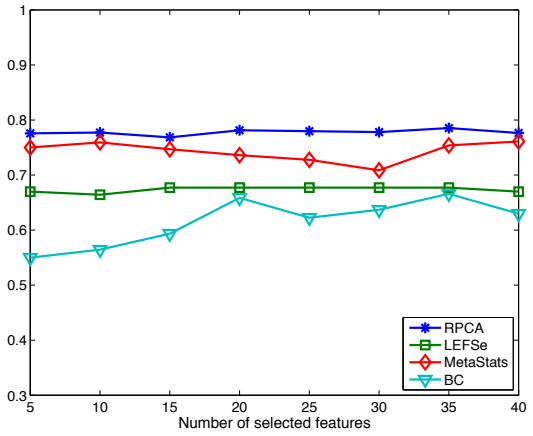

Fig. 2. Balanced accuracy of the four algorithms using NCC-L1 classification over the canine IBD dataset.

for the two classes accounts for preventing the accuracy to be dominated by the class with the majority samples. If the dataset is balanced, the balanced accuracy decreases to the conventional accuracy.

Furthermore, we used the same evaluation procedure to confront our method with the latest bacterial biomarker detection schemes: LEFSe, MetaStats, and binary classification (BC) [22]. The result of the comparisons over the two datasets using two classification methods are presented in Figures 1-4. The results illustrate that our method significantly outperformed the existing approaches on the presented test cases. Our method yielded consistent classification accuracy across both datasets and for both classification methods. Furthermore, our method showed consistency when the number of selected features is varied while showing only mild decrease in classification accuracy when the number of selected features is significantly decreased. It is worth mentioning that the classification accuracy is expected to decrease when the number of selected features decreases due to the loss of potentially important classification information from the dropped features.

Figures 5 and 6 present the log-base-10 of the top ten bacterial biomarkers for IBD produced by RPCA on both the canine and mouse datasets, respectively. These bacterial biomarkers exhibited substantial abundance differentiation between healthy and diseased samples. Identifying bacterial biomarkers for IBD enables further investigation of these bacteria to understand their link to IBD.

IV. CONCLUSION

This paper proposed and exploited a new approach to identify microbial biomarkers for inflammatory bowel disease. IBD can cause severe discomfort and in some cases could lead to life-threatening complications. Identifying bacterial biomarkers for IBD enables the detection and a deeper understanding of IBD. We divide the bacterial biomarker identification problem into two steps. In the first step we apply RPCA to separate the bacteria exhibiting differential abundance between

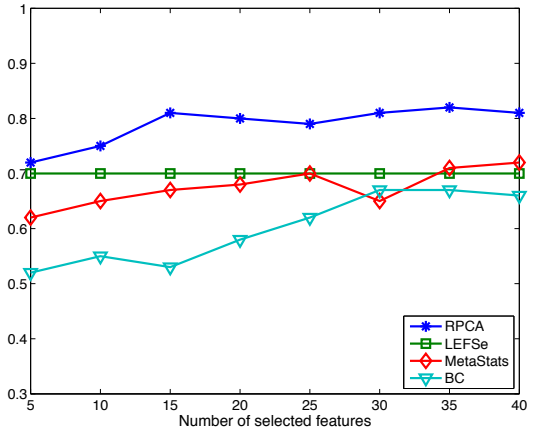

Fig. 3. Balanced accuracy of the four algorithms using kNN classification over the mouse model of UC dataset.

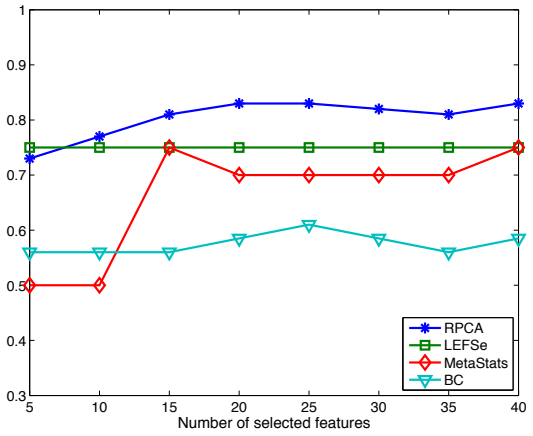

Fig. 4. Balanced accuracy of the four algorithms using NCC-L1 classification over the mouse model of UC dataset.

diseased and control samples from the bacteria with non-differential abundance. The second step employs a scoring

ISBN 978-0-9928626-7-1 © EURASIP 2017

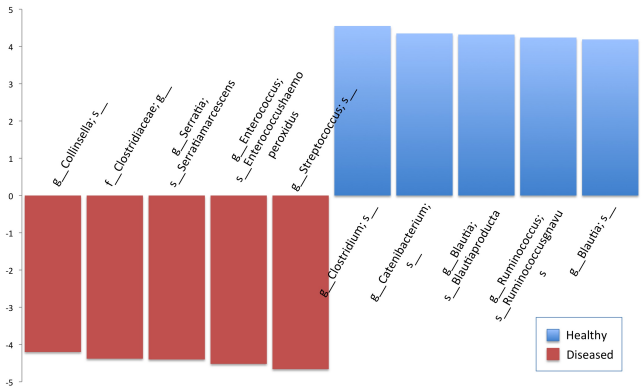

Fig. 5. The log-base-10 RPCA score of the top ten bacteria biomarkers for IBD in the canine dataset as outputted by RPCA.

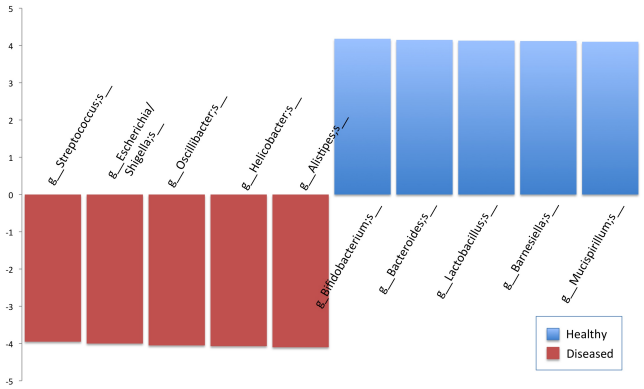

Fig. 6. The log-base-10 RPCA score of the top ten bacteria biomarkers for UC in the mouse dataset as outputted by RPCA.

process on the differentially abundant bacteria to identify the top bacteria to be selected as biomarkers. We compared our method with three well used bacterial biomarker detection methods over two datasets and using two classification methods. Our method significantly outperformed the other methods over the different evaluation cases.

Acknowledgment

The work of M. Alshawaqfeh and E. Serpedin was supported by an research gift made available by Ooredoo.

References

[1] S. B. Hanauer, "Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities," *Inflammatory Bowel Diseases*, vol. 12, no. 5, pp. S3–S9, 2006.

[2] 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.

[3] A. B. Shreiner, J. Y. Kao, and V. B. Young, "The gut microbiome in health and in disease," *Current Opinion in Gastroenterology*, vol. 31, no. 1, p. 69, 2015.

[4] B. Khor, A. Gardet, and R. J. Xavier, "Genetics and pathogenesis of inflammatory bowel disease," *Nature*, vol. 474, no. 7351, pp. 307–317, 2011.

[5] 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.

[6] 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.

[7] 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.

[8] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[9] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, vol. 3, 2003, pp. 856–863.

[10] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[11] J. Yang, Y. Liu, Z. Liu, X. Zhu, and X. Zhang, "A new feature selection algorithm based on binomial hypothesis testing for spam filtering," *Knowledge-Based Systems*, vol. 24, no. 6, pp. 904–914, 2011.

[12] D. Huang and S. Meyn, "Feature selection for composite hypothesis testing with small samples: Fundamental limits and algorithms," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1917–1920.

[13] K. Torkkola, "Feature extraction by non parametric mutual information maximization," *The Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.

[14] W. Duch, J. Biesiada, T. Winiarski, K. Grudziński, and K. Grabczewski, "Feature selection based on information theory filters," in *Neural Networks and Soft Computing*. Springer, 2003, pp. 173–178.

[15] 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.

[16] 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.

[17] 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*, 2009, pp. 2080–2088.

[18] 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.

[19] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[20] H. Park, M. Jeon, and J. B. Rosen, "Lower dimensional representation of text data based on centroids and least squares," *BIT Numerical Mathematics*, vol. 43, no. 2, pp. 427–448, 2003.

[21] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *The 20th International Conference on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 3121–3124.

[22] 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.