Advances in
Biotechnology & Microbiology
ISSN 2474-7637

Juniper Publishers
key to the Researchers

Mini Review
Volume 6 Issue 4 – October 2017
DOI: 10.19080/AIBM.2017.06.555695

Adv Biotech & Micro

Machine Learning Techniques on Microbiome -Based Diagnostics

Lihong LI and Damian R Mingle*

Intermedix, USA

Submission: August 31, 2017**; Published:** October 20, 2017

***Corresponding author:** Damian R Mingle, Chief Data Scientist, Intermedix, Nashville TN 37219, USA, Tel: 615-364-9660; Email: Damian.Mingle@Intermedix.com

Introduction

Compared with our 10 trillion human cells, the human microbiota consists of the 10-100 trillion microbial cells, with the vast majority living in the human gastrointestinal tract [1]. Microbial communities play an essential role to human health. For example, the gut microbial symbioses contribute to its host to perform a multitude of functions, including, but not limited to, digestion and production of nutrients, detoxification, protection against pathogens and regulation of immune system [2]. However, the symbiosis of microbial communities can cause disease. For example, alterations of gut microbial communities can lead to autoimmune disorders, vaginal communities to bacterial vaginosis. Therefore, identifying important features of the microbial community or finding the association between the composition of microbiota and clinical features of disease will improve detection, diagnosis and therapeutic monitoring of disease.

With the aid of rapidly developing, next-generation sequencing methods and computational approaches used to maximally extract meaningful patterns from the high-dimensional human microbiota surveys data, researchers can now enhance the ability to understand the composition of human microbiome. Applying machine learning and statistical techniques in human microbiome data to address the complex mechanisms underlying disease has brought about a new field of macrobiotics. Sequencing of the 16S rRNA gene is an effective method for interrogating the taxonomic composition of microbial communities [3]. This gene is used as the standard for classification and identification of microbes since it presents in most microbes and shows proper changes.

Machine Learning on Microbiome-Based Diagnostics

The 16S rRNA is commonly used in human microbiota survey studies. Operational taxonomic units (OTUs), typically based on similar 16S rRNA sequences, have been the most commonly used units of microbial diversity. Therefore, abundances of OTUs from 16S rRNA gene sequence surveys can serve as useful input features for classification problems in macrobiotic data [4]. Prior work by Knights et al. [4] took an excellent first step in establishing the feasibility of creating accurate models for classification of body sites and subject identification. Because of the generally large number of features in microbiota classification, they demonstrated three categories in feature selection: filter methods, wrapper methods and embedded methods.

They further discussed mechanism learning algorithm on five benchmark classification tasks data from bacterial 16S rRNA gene-based surveys of various human body habitats. This discussion led to a realization that that random forest (RF) was the strongest performer and multinomial naïve Bays also performed well. Although elastic net (ENET) classifier had noticeably higher expected error than RF, it still proved useful for performing feature selection as a preprocessing step for other classifiers. Statnikov et al. [5] performed a systematic comparison of 18 major machine learning methods in seven algorithmic families (support vector machines, kernel ridge regression, regularized logistic regression, Bayesian logistic regression, random forests, k-nearest neighbors and probabilistic neural networks) for multi category classification, 5 feature/OTU selection methods, using 8 datasets with 16S rRNA gene surveys spanning 1,802 human samples and various classification tasks: body site and subject classification and diagnosis. They found that RFs, support vector machines, kernel ridge regression and Bayesian logistic regression with Laplace priors are the most effective machine learning techniques for performing accurate classification on these macrobiotic data.

The machine learning study of microbial compositions yielded important clues for understanding, diagnosing and treating disease by inferring the contribution of each constituent

of microbiota to various disease and physiological states [5]. Inflammatory bowel disease (IBD), an autoimmune condition, is observed to be associated with major alterations in the gut microbiome taxonomic composition. Papa et al. [6] investigated the microbiome-based diagnostics with 16S rRNA sequencing of fecal samples by machine learning methods to distinguish pediatric patients with IBD from patients with similar symptoms. Basically, they assigned each sequence in the data set to a taxonomical group using Naïve Bayesian classifier. For each sample they then calculated the relative abundance of each tax with respect to the total number of sequences in each sample. Finally, they then trained a RF classifier to assign the class (IBD or non-IBD) based on the relative sequence abundances in every tax. The method identified disease-associated tax and distinguished patients with Crohn's disease from those with ulcerative colitis with reasonable accuracy. It showed that classification based on microbial diversity is an effective complementary technique for IBD detection in pediatric patients.

Besides the 16S rRNA-based studies considering the taxonomic structure of microbial communities, Yazdani et al. [7] focused on understanding the functional profiles of IBD microbiomes to determine how microbial changes in the health and disease status function. To classify major changes in microbiome protein family abundances between healthy subjects and IBD patients, they applied machine learning on results obtained previously from computing relative abundance of ~10,000 Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologous protein families in the gut microbiome of a set of healthy individuals and IBD patients. They developed and trained a two-step classifier for identifying major shifts in human gut microbiome protein family abundance between healthy and IBD cohorts with the Kolmogorov-Smirnov (KS) test and RFs.

Irritable bowel syndrome (IBS) is a functional gastrointestinal disorder involving multiple path physiological mechanisms in which composition of gut microbiota has been proposed as one of the potentially important factors. To identify the association between the composition of the intestinal microbiota and clinical features of IBS, Tap et al. [8] implemented L1 regularized logistic regression (least absolute shrinkage and selection operator (LASSO)) on collected fecal and mucosal samples from IBS patients and healthy subjects. The LASSO procedure identified 90 bacterial OTUs that could be used as a composite gut microbial signature for IBS severity. Saulnier et al. [9] applied the random forest (RF) approach to analyze 71 samples from 22 children with IBS and 22 healthy children. With random forest techniques, they were able to classify different subtypes of IBS with a success rate of 98.5%. Their findings indicated the important association between gastrointestinal microbes and IBS in children.

Bacterial Vaginosis (BV) is a disease associated with the vagina microbiome and caused by an imbalance of the naturally occurring bacteria in the vagina. However, it is difficult to identify a single cause of BV even though the microbial community and BV are correlated. To identify important features of the microbial community, Beck et al. [2] employed three machine learning techniques: genetic programming (GP), RFs and logistic regression (LR). This study demonstrates the feasibility of using classification models to identify populations in a microbial community that are associated with BV, with accuracies above 90% for Nugent score BV and above 80% for Amsel criteria BV obtained by the classification models. Carter et al. [10] successfully applied GEFES (Genetic & Evolutionary Feature Selection), genetic algorithm (GA) based feature selection, to identify the key features in the human vaginal microbiome and in patient meta-data that are associated with BV. The dataset used for their experiment consisted of 1601 instances with 410 features, including relative abundances of bacterial species determined by next-generation sequencing of 16S rRNA fingerprint sequences from 25 women over a 10 week period. It showed that GA-based feature selection can increase classification accuracy of BV from microbiome data using fewer features.

The machine learning approach also provided an opportunity to improve the sensitivity of noninvasive tests to identify shifts in the composition of the gut microbiota associated with the progression of colorectal cancer (CRC), the second leading cause of death among cancers in the United States. Baxter et al. [11] developed a random forest model with the relative abundances of the bacterial populations within each sample to detect colonic lesions, based on 16S rRNA genes from the stool samples of 490 patients. The microbiota-based RF model detected 91.7% of cancers and 45.5% of adenomas, whereas widely used fecal immunochemical test alone detected 75.0% and 15.7%. These findings demonstrate the potential for microbiota analysis to complement existing screening methods for improved detection of colonic lesions [12].

Conclusion

In this paper we have provided a brief overview of the machine learning approach and its usefulness in terms of various microbiome-based diagnostics. A number of effective machine learning applications have demonstrated their potential on analysis of complex microbiota communities, and it is expected that research will continue to improve upon and extend these techniques. The use of machine learning has allowed us to better explore the microbial data and provide new insights on the interactions between microbial factors and their effects on the host. Moving forward, the machine learning approaches are believed to constitute credible starting points for further research on microbiome-based diagnostics to identify specific disease-associated microbial communities.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser Liggett CM, Knight R, et al. (2007) The human microbiome project. Nature 449(7164): 804-810.

How to cite this article: Lihong LI, Damian R M. Machine Learning Techniques on Microbiome-Based Diagnostics. Adv Biotech & Micro. 2017; 6(4): 555695. DOI: 10.19080/AIBM.2017.06.555695.

2. Beck D, Foster JA (2014) Machine learning techniques accurately classifies microbial communities by bacterial vaginosis characteristics. PloS one 9(2): e87830.

3. Jung Wu H, Wu E (2012) The role of gut microbiota in immune homeostasis and autoimmunity. Gut microbes 3(1): 4-14.

4. Knights D, Costello EK, Knight R (2011) Supervised classification of human microbiota. FEMS Microbiol Rev 35(2): 343-359.

5. Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, et al. (2013) A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome 1: 11.

6. Papa E, Docktor M, Smillie C, Weber S, Preheim SP, et al. (2012) Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. PloS one 7(6): e39242.

7. Yazdani M, Taylor BC, Debelius JW, Li W, Knight R, et al. (2016) Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease. In Big Data (Big Data), 2016 IEEE International Conference. pp. 1272-1280.

8. Tap J, Derrien M, Törnblom H, Brazeilles R, Cools Portier S, et al. (2017) Identification of an intestinal microbiota signature associated with severity of irritable bowel syndrome. Gastroenterology 152(1): 111-123.

9. Saulnier DM, Riehle K, Mistretta TA, Diaz MA, Mandal D, et al. (2011) Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. Gastroenterology 141(5): 17820-1791.

10. Carter J, Beck D, Williams H, Foster J, Dozier G, et al. (2014) GA-Based selection of vaginal microbiome features associated with bacterial vaginosis. Genet Evol Comput Conf 2014: 265-268.

11. Baxter NT, Ruffin MT, Rogers MAM, Schloss PD (2016) Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome medicine 8: 37.

12. Gevers D, Pop M, Schloss PD, Huttenhower C (2012) Bioinformatics for the human microbiome project. PLoS Comput Biol 8(11): e1002779.

DOI: 10.19080/AIBM.2017.06.555695

How to cite this article: Lihong LI, Damian R M. Machine Learning Techniques on Microbiome-Based Diagnostics. Adv Biotech & Micro. 2017; 6(4): 555695. DOI: 10.19080/AIBM.2017.06.555695.