Microbiomarkers Discovery in Inflammatory Bowel Diseases using Network-Based Feature Selection

Mostafa Abbas
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
mohamza@hbku.edu.qa

Thanh Le
College of Information Sciences and Technology
Pennsylvania State University
University Park, PA 16802
txl252@ist.psu.edu

Halima Bensmail
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
hbensmail@hbku.edu.qa

Vasant Honavar
College of Information Sciences and Technology
Pennsylvania State University
University Park, PA 16802
vhonavar@ist.psu.edu

Yasser EL-Manzalawy
College of Information Sciences and Technology
Pennsylvania State University
University Park, PA 16802
yme2@psu.edu

ABSTRACT

Discovery of disease biomarkers is a key step in translating advances in genomics into clinical practice. There is growing evidence that changes in gut microbial composition are associated with the onset and progression of Type 2 Diabetes (T2D), Obesity, and Inflammatory Bowel Disease (IBD). Reliable identification of the most informative features (i.e., microbes) for discriminating metagenomics samples from two or more groups (i.e., phenotypes) is a major challenge in computational metagenomics. We propose a Network-Based Biomarker Discovery (NBBD) framework for detecting disease biomarkers from metagenomics data. NBBD has two major customizable modules: i) A ***network inference module*** for inferring ecological networks from the abundances of microbial operational taxonomic units (OTUs); ii) A ***node importance scoring module*** for comparing the constructed networks for the chosen phenotypes and assigning a score to each node based on the degree to which the topological properties of the node differ across two networks. We empirically evaluated the proposed NBBD framework, using five network inference methods for inferring gut microbial networks combined with six node topological properties, on the identification of IBD biomarkers using a large dataset from a cohort of 657 and 316 IBD and healthy controls metagenomic biopsy samples, respectively. Our results show that NBBD is very competitive with some of the state-of-the-art feature selection methods including the widely used method based on random forest variable importance scores.

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233602

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; • **Applied computing**→ Biological networks; Computational genomics

KEYWORDS

Metagenomics, Inflammatory Bowel Diseases, Biological networks analysis, Feature selection

ACM Reference format:

Mostafa Abbas, Thanh Le, Halima Bensmail, Vasant Honavar and Yasser EL-Manzalawy. 2018. Microbiomarkers Discovery in Inflammatory Bowel Diseases using Network-Based Feature Selection. In ***Proceedings of ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB'18). ACM, Washington, DC, USA, 6 pages.*** https://doi.org/10.1145/3233547.3233602

1 Introduction

Inflammatory bowel disease (IBD) is a group of disorders that is characterized by flares of inflammation in the gut. Several studies have shown that the gut microbiota plays an important role in the pathogenesis of IBD [10; 15; 16; 20]. Recent advances in sequencing technology have expanded rapidly the amount of metagenomics samples collected from the gut under different health/disease conditions [4; 32]. Despite the existence of several IBD metagenomics datasets (e.g., [10; 13; 34]) and metagenome-wide analysis studies (e.g. [6; 10; 13]), the role of the gut

microbiome in the pathogenesis of IBD remains poorly understood [10]. Therefore, there is an urgent need for effective methods for data analysis, interpretation, and translation of the resulting insights into clinical practice [30]. Of particular interest are computational and statistical methods for integrative analyses of large metagenomics datasets to discover reliable biomarkers of IBD disease [29] as well as microbial signatures for different IBD subtypes [23].

Identification of disease biomarkers from metagenomics data calls for effective methods for selecting, from a very large number of candidate features, a small subset of features that can accurately discriminate between the phenotypes (e.g., IBD versus healthy). This task is very challenging in practice due to [28]: i) curse of dimensionality (i.e., large number of features and small numbers of samples); ii) high degree of sparsity of the metagenomics data samples (where only a small fraction of the entries have non-zero values); iii) complexity of the underlying biology and limitations in sequencing technology and taxonomy classification pipelines. To address these challenges, several statistical methods have been proposed in the literature to compare abundance of features (e.g., genes or OTUs) between two groups [36]. Some of these methods have been designed specifically for RNA-Seq data (e.g., DESeq [1] and edgeR [27]) while recently tools such as metagenomeSeq [24] and analysis of composition of microbiomes (ANCOM) [19] have been developed specifically for metagenomics data, which is often more sparse than RNA-Seq data. Machine learning based feature selection [11] is another widely used approach for identifying the most discriminative (informative) features from either RNA-Seq or metagenomics data.

Against this background, we present an integrative framework for Network-Based Biomarkers Discovery (NBBD). NBBD integrates comparative network analysis for prioritizing biomarkers and the machine learning approach for assessing the discriminative power of the top selected biomarkers. We tested the proposed framework on the challenging task of identifying biomarkers from a large dataset of new-onset IBD metagenomics biopsy samples collected from pediatrics. Using our framework as a test-bed for evaluating five commonly used ecological network inference tools and six node topological properties, our results suggest that networks inferred from the same data but using different tools have substantial differences in their topological properties. Moreover, our method can identify highly discriminative biomarkers even from poorly inferred networks (e.g., networks with high rates of false positive and/or negative edges). Our results also suggest that the network-based feature selection method is very competitive with some state-of-the-art feature selection methods for determining the most discriminative features from metagenomics data. Finally, analyses of the identified IBD biomarkers suggest promising candidates for targeted experimental studies.

2 Materials and Methods

2.1 Datasets

The OTU BIOM files and meta-data (including age, gender, race, disease severity, behavior, and location) for a large cohort IBD dataset [10] were downloaded from QIITA (https://qiita.ucsd.edu/) database. The dataset consists of 1359 metagenomics samples including rectal tissue biopsy and fecal samples. We filtered the dataset by discarding fecal samples and samples corresponding to patients with age greater than 18 years. Thus, our final dataset consists of 657 and 316 IBD and healthy control metagenomic biopsy samples, respectively. We then randomly split the final dataset into training and test datasets such that the training data has 200 IBD and 200 healthy samples. Each sample has 786 OTUs at the genus level that were extracted using summarize_taxa.py QIIME script.

2.1 Network-based Biomarker Discovery (NBBD) Framework

NBDD framework consists of two main customizable modules, a network inference module and a node importance scoring module. Fig. 1 provides an overview of the NBBD framework: Given a pair of OTU tables (e.g., corresponding to IBD and healthy samples), the network inference module constructs a microbial network from each OTU table. In these networks, each node corresponds to an OTU and each edge represents a relationship between two nodes (e.g., co-occurrence). The node importance scoring module compares the two networks and assigns a score to each node based on the degree to which the topological properties of the node differ across the two networks. We hypothesize that the nodes that show the greatest difference across the two networks should provide useful features for training a classifier to discriminate between two populations of metagenomics samples.

Let $G_i(V_i, E_i)$ and $G_j(V_j, E_j)$ represent two graphs (networks) constructed from two groups (i, j) of metagenomics samples. We score each node $v \in V_i \cap V_j$ with respect to a node property P as:

$$score^P(v) = |f_P(v, G_i) - f_P(v, G_j)|$$

where $f_P(v, G_i)$ is the value of the property $\boldsymbol{P}$ for node v in G_i. For instance, $f_P(v, G_i)$ could be the degree of v in G_i.

2.3 Network Inference Methods

We experimented with five widely used microbial network inference methods. We used the default parameters of each tool, unless noted otherwise. In what follows, we briefly summarize each of the methods.

SparCC: Sparse Correlations for Compositional data (SparCC) [8] infers a network of associations between the microbial species based on the linear Pearson correlation between the log-transformed OTUs, under the assumption that the underlying network is sparse. We used the implementation of SparCC provide as part of SPIEC-EASI [17] tool.

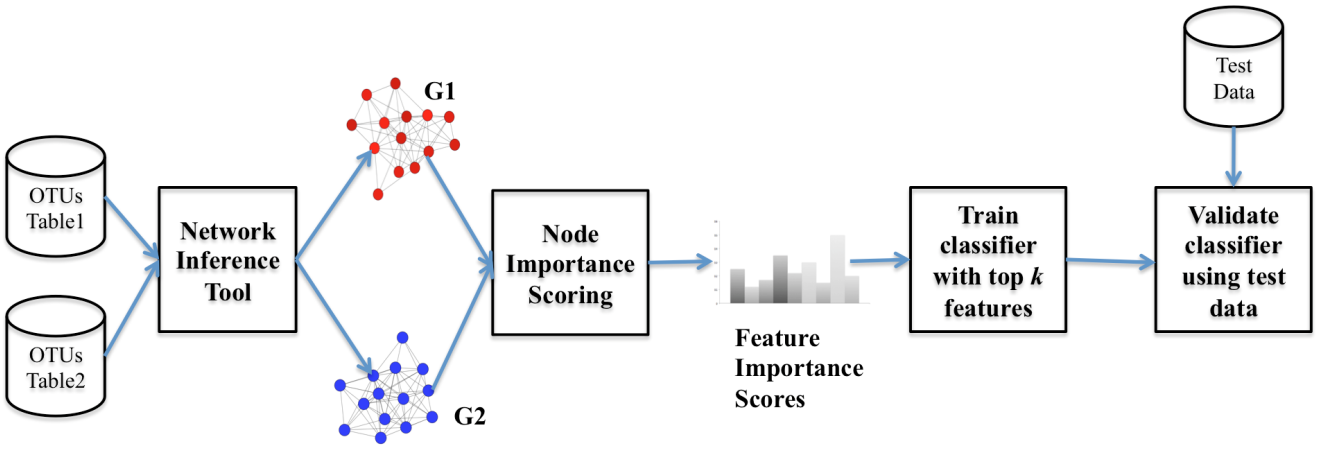

Figure 1: Overview of the NBBD framework. Training data in the form of two OTU tables corresponding to two groups of metagenomics samples are first used to construct two networks. The node importance scoring modules compares topological properties of shared nodes in the two graphs and outputs scores to prioritize the input features. Top selected features are then used to train and evaluate a classifier.

Glasso: Graphical lasso (Glasso) [9] estimates a network of associations between OTUs by estimating a sparse inverse of the covariance matrix. Its advantages include speed and the reliance on only one parameter to be tuned (the regularization parameter which controls the sparsity of the learned network). We used the Glasso implementation that is part of SPIEC-EASI [17].

MB: This method, due to Meinshausen and Bühlmann [20] (hence the name MB method), estimates sparse networks by identifying direct neighbors (for each node) as the smallest subset of nodes such that the target node is conditionally independent of the rest of the networks given the direct neighbors so identified. MB is also implemented in SPIEC-EASI [17].

RMT: This method uses Pearson correlation coefficient to add an edge between two OTUs if their correlation is higher than a threshold that is optimized using a procedure based on the Random Matrix Theory (RMT). The method is implemented in the Molecular Ecological Network Analysis Pipeline [5]. We used the default parameters except for the parameter controlling the number of OTUs that build the network. An OTU was used if it is expressed in at least 25% of the samples. The default setting of 50% fails to construct a network.

CoNet: This method infers the association network by combining two complementary approaches [7]: an ensemble method of similarity or dissimilarity measures; and a novel permutation-renormalization bootstrap method, ReBoot [7], to assess the significance of the associations.

2.4 Node Topological Properties

Let G (V, E) be a network (or graph) where V and E denote the sets of nodes and edges, respectively. We considered the following node properties implemented in NetworkX [12]:

Betweenness Centrality (btw): Betweenness centrality of a node v is determined as $f_{btw}(v,G) = \sum_{u \neq w \in V} \frac{\sigma(u,w|v)}{\sigma(u,w)}$ where $\sigma(u,w)$ is the total number of shortest paths between u and w, and $\sigma(u,w|v)$ is the number of shortest paths between u and w passing through v.

Closeness Centrality (cls): Closeness centrality of a node v is $f_{cls}(v,G) = \frac{n-1}{\sum_{u=1}^{n-1} d(u,v)}$ where $d(u,v)$ is the shortest path distance between u and v, and n is the number of nodes that can reach v.

Average Neighbor Degree (and): The average neighborhood degree of a node v is $f_{and}(v,G) = \frac{1}{|N(v)|}\sum_{u \in N(v)} k_u$ where $N(v)$ are the neighbors of node v and k_u is the degree of node $u \in N(v)$.

Clustering Coefficient (cc): For unweighted graphs, the clustering coefficient of a node v is $f_{cc}(v,G) = \frac{2T(v)}{\deg(v)(\deg(v)-1)}$ where $T(v)$ is the number of triangles that include node v and $\deg(v)$ is the degree of v.

Node Clique Number (ncn): The node clique number of a node v is the size of the largest maximal clique containing v, where a clique is a subset of nodes such that there is an edge between every pair of distinct nodes.

Core Number (cn): The core number of a node v is the largest value k of a k-core containing v, where k-core is a maximal subgraph that contains nodes of degree k or more.

2.5 Machine Learning Classifiers and Performance Evaluation

We used the training data to train Random Forest (RF) [3] classifiers to discriminate between (positively labeled) IBD

samples and (negatively labeled) healthy samples. We used the implementation of RF algorithm provided in Scikit-learn [25] and set the number of trees to 500. We evaluated the performance of the resulting classifiers on the test set using a set of commonly used performance measures: Accuracy (ACC), Sensitivity (Sn), Specificity (Sp), Mathew's Correlation Coefficients (MCC), and Area Under ROC Curve (AUC) [2].

3 Results and Discussion

3.1 Exploratory Analysis

We used PICRUSt [18] to examine the functional space in the IBD and healthy microbiome samples. PICRUSt infers functional activity by constructing ancestral gene content and then estimating the abundance of gene families in the 16S rRNA. The resultant functional gene count matrix was first normalized to rescale counts per sample to lie in the interval [0,1] before conducing follow-up analyses using Principal component analysis (PCA) and Student's ***t***-test. Fig. 2-a and 2-b show the visualization of the first two principal components when PCA was applied to functional and compositional profiles of our training data, respectively. We did not see evidence of two distinct groups corresponding to IBD and healthy data in PC space. This led us to conjecture that non-linear classifiers (e.g., Random Forests [3]) will outperform linear models (e.g., Support Vector Machine [33] with a linear kernel) on this data. The statistically significant (at a p-value < 0.05) functional differences between IBD and healthy samples are shown in Fig. 2-c. The results suggest that the IBD group exhibits a decrease with respect to 15 KEGG metabolic pathways, many of which have been reported in the literature. For example, decreased level of Tryptophan in serum was shown to be significantly lower in IBD patients [22], which is consistent with the lower microbiome Tryptophan metabolism activity in our IBD samples relative to the healthy samples. Medicherla et al. [21] have shown that the oral administration of geraniol inhibit pro-inflammatory cytokines in patients with murine colitis, which is again consistent with what we observe in our samples. A lower geraniol degradation activity in the IBD samples suggests lower availability of geraniol.

The top 10 most differentially abundant OTUs between IBD and healthy samples identified using the nonparametric Kruskal-Wallis statistical test [14] are shown in Fig. 3. Surprisingly, only six OTUs have significant (adjusted p-values < 0.05) differential abundance with respect IBD and healthy groups. These OTUs correspond to Clostridiaceae and Pasteurellaceae families and four genera, Blautia, Coprococcus, Roseburia, and Ruminococcus.

3.2 Feature Selection Improves the Predictive Performance of RF Classifiers

Table 1 compares the performance on the test set of a RF classifier trained using all 786 OTUs and RF classifiers trained using top 30 OTUs determined from training data using the following commonly used feature selection methods, RF feature importance [3], Lasso [31], Information Gain (IG), and Min- Redundancy and

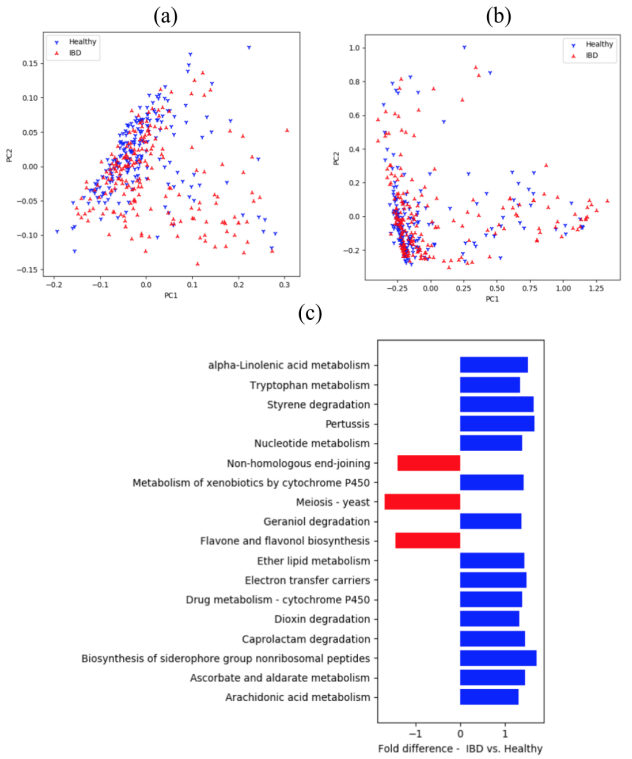

Figure 2: Visualization of the first two Principal components from the PCA analysis of (a) PICURSt functional profiles and (b) normalized OTU counts. (c) Functional differences, predicted using PICRUSt, of statistically significant KEGG metabolic pathways.

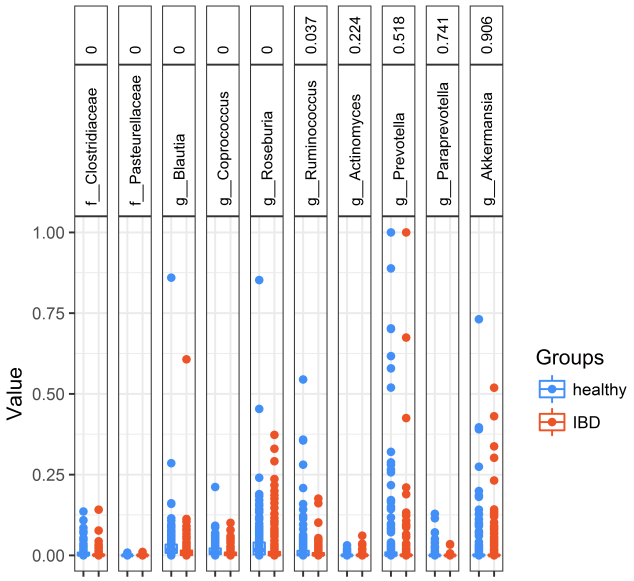

Figure 3: Top 10 most differentially abundant OTUs between IBD and healthy samples identified using the nonparametric Kruskal-Wallis statistical test.

Table 1: Performance of RF classifiers trained using all features and top 30 features selected using different feature selection methods.

Feature Selection	ACC (%)	Sn	Sp	MCC	AUC
None	66	0.64	0.75	0.31	0.74
RF	66	0.64	0.77	0.33	0.78
Lasso	41	0.34	0.68	0.02	0.52
IG	65	0.61	0.81	0.34	0.75
MRMR	43	0.38	0.64	0.02	0.52

Max-Relevance (MRMR) [26] using F-Statistic and Pearson's correlation coefficient for assessing relevance of and redundancy between features. Using all features, the performance (in terms of AUC) of the RF classifier estimated using the test data is 0.74. On the other hand, a RF classifier trained using top 30 features, determined using feature importance of another trained RF classifier, had AUC score of 0.78. Surprisingly, classifiers trained using top 30 features determined using Lasso or MRMR methods have very poor performance. One possible explanation of this finding is that the basic Lasso method fits a linear model whereas as suggested by our exploratory analyses, IBD and healthy samples cannot be reliably discriminated using a linear model. In the case of MRMR, our results seem to suggest that F-Statistic and/or Pearson's Correlation Coefficient do not reliably estimate the relevance and/or redundancy when the feature space is extremely sparse.

3.3 Performance of Network-based Feature Selection Methods

Table 2 reports the best performing classifier (in terms of AUC) using NBBD feature selection for each choice of the network inference methods. The highest AUC of 0.77 is obtained using SparCC combined with node betweenness centrality (btw) for determining node importance scores or RMT combined with core number (cn) node property for computing node importance scores. These results are especially noteworthy in light of a recently published comparative study [35] which showed that SparCC, RMT, and CoNet (among other correlation network inference tools) suffer from extremely poor precision (i.e., below 0.2), and hence yield networks with a large number of false edges. Our results show that even networks with a large fraction of spurious edges can be used to reliably identify potential disease markers from metagenomic data.

3.4 Analysis of Identified IBD Microbial Markers

Tables 1 and 2 show that there exist three classifiers (each trained using a subset of top 30 selected OTUs) with AUC scores in the range 0.77-0.78. We named the corresponding feature subsets according to the feature selection method used to identify them (e.g., SparCC_btw, RMT_cn, and RF). We noted that the combination of these 90 OTUs resulted in 50 unique OTUs.

Table 2: Performance of top performing RF classifier (in terms of AUC) for each network inference method in NBBD.

Tool	Property	ACC (%)	Sn	Sp	MCC	AUC
SparCC	btw	66	0.62	0.79	0.34	0.77
Glasso	btw	58	0.57	0.59	0.14	0.63
MB	cc	57	0.53	0.73	0.21	0.66
RMT	cn	66	0.63	0.78	0.33	0.77
CoNet	btw	63	0.60	0.78	0.30	0.74

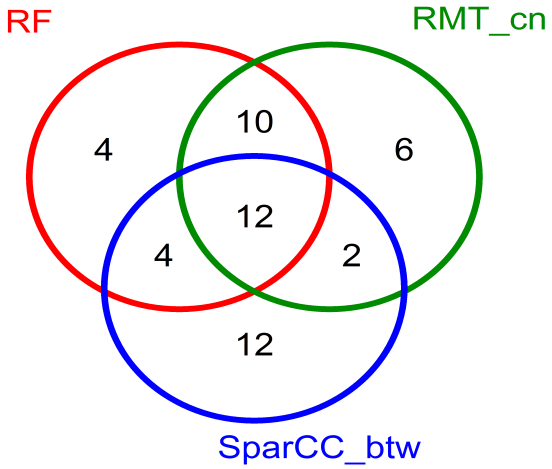

Figure 4: Venn diagram of unique and shared features selected using RF feature importance, network-based feature selection applied to RMT (SparCC) networks and using cn (btw) for node importance scoring.

The Venn diagram of unique and shared OTUs among the three subsets of features is given in Fig. 4. The number of unique OTUs in each subset is 12, 6, and 4 for SparCC_btw, RMT_cn, and RF sets, respectively. The network-based feature selection method that shares the largest number (22) of OTUs in common with those identified using RF feature importance method is RMT_cn. We further observed that there were only 12 OTUs shared among the three sets. To assess the significance of the difference between the medians of relative abundance of these OTUs in IBD and healthy populations, we applied the Mann-Whitney nonparametric test. We found lower (but non statistically significant) abundance of Faecalibacterium genus in IBD samples. We also found significantly higher abundance of Gemellaceae and Sutterella in the IBD samples relative to the healthy samples. In the case of the remaining nine OTUs (Parabacteroides, Clostridiales, Clostridiaceae, Ruminococcus, Coprococcus, Lachnospira, Roseburia, Erysipelotrichaceae, Eubacterium), our results show significantly lower abundances in IBD samples relative to the healthy samples. Mechanistic understanding of the precise reasons for these observed differences calls for controlled experiments.

4 Conclusions

We proposed a novel Network-Based Biomarker Discovery (NBBD) framework for detecting disease biomarkers from metagenomics data. NBBD consists of: a network inference module, for inferring networks from the abundances of microbial operational taxonomic units (OTUs); and a node importance scoring module, for comparing the constructed networks for the chosen phenotypes and assigning a score to each node based on the degree to which the topological properties of the node differ across constructed networks. Our results show that the NBBD approach is able to reliably identify IBD biomarkers even when the constructed networks have high rates of false positive edges.

ACKNOWLEDGMENTS

Research supported in part by the Center for Big Data Analytics and Discovery Informatics at the Pennsylvania State University and by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1 TR000127 and TR002014. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

[1] Anders, S. and Huber, W., 2010. Differential expression analysis for sequence count data. ***Genome biology 11***, 10, R106.

[2] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., and Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. ***Bioinformatics 16***, 5, 412-424.

[3] Breiman, L., 2001. Random forests. ***Machine learning 45***, 1, 5-32.

[4] Debelius, J.W., Vázquez-Baeza, Y., Mcdonald, D., Xu, Z., Wolfe, E., and Knight, R., 2016. Turning participatory microbiome research into usable data: lessons from the American Gut Project. ***Journal of microbiology & biology education 17***, 1, 46.

[5] Deng, Y., Jiang, Y.-H., Yang, Y., He, Z., Luo, F., and Zhou, J., 2012. Molecular ecological network analyses. ***BMC bioinformatics 13***, 1, 113.

[6] Eck, A., De Groot, E., De Meij, T., Welling, M., Savelkoul, P., and Budding, A., 2017. Robust microbiota-based diagnostics for inflammatory bowel disease. ***Journal of clinical microbiology 55***, 6, 1720-1732.

[7] Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C., 2012. Microbial co-occurrence relationships in the human microbiome. ***PLoS computational biology 8***, 7, e1002606.

[8] Friedman, J. and Alm, E.J., 2012. Inferring correlation networks from genomic survey data. ***PLoS computational biology 8***, 9, e1002687.

[9] Friedman, J., Hastie, T., and Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. ***Biostatistics 9***, 3, 432-441.

[10] Gevers, D., Kugathasan, S., Denson, L.A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S.J., and Yassour, M., 2014. The treatment-naive microbiome in new-onset Crohn's disease. ***Cell host & microbe 15***, 3, 382-392.

[11] Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. ***Journal of machine learning research 3***, Mar, 1157-1182.

[12] Hagberg, A., Swart, P., and S Chult, D., 2008. ***Exploring network structure, dynamics, and function using NetworkX.*** Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[13] Halfvarson, J., Brislawn, C.J., Lamendella, R., Vázquez-Baeza, Y., Walters, W.A., Bramer, L.M., D'amato, M., Bonfiglio, F., Mcdonald, D., and Gonzalez, A., 2017. Dynamics of the human gut microbiome in inflammatory bowel disease. ***Nature microbiology 2***, 5, 17004.

[14] Hollander, M., Wolfe, D.A., and Chicken, E., 2013. ***Nonparametric statistical methods***. John Wiley & Sons.

[15] Kamada, N., Seo, S.-U., Chen, G.Y., and Núñez, G., 2013. Role of the gut microbiota in immunity and inflammatory disease. ***Nature Reviews Immunology 13***, 5, 321.

[16] Kostic, A.D., Xavier, R.J., and Gevers, D., 2014. The microbiome in inflammatory bowel disease: current status and the future ahead. ***Gastroenterology 146***, 6, 1489-1499.

[17] Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., and Bonneau, R.A., 2015. Sparse and compositionally robust inference of microbial ecological networks. ***PLoS computational biology 11***, 5, e1004226.

[18] Langille, M.G., Zaneveld, J., Caporaso, J.G., Mcdonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Thurber, R.L.V., and Knight, R., 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. ***Nature biotechnology 31***, 9, 814.

[19] Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., and Peddada, S.D., 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. ***Microbial ecology in health and disease 26***, 1, 27663.

[20] Manichanh, C., Reeder, J., Gibert, P., Varela, E., Llopis, M., Antolin, M., Guigo, R., Knight, R., and Guarner, F., 2010. Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. ***Genome research 20***, 10, 1411-1419.

[21] Medicherla, K., Sahu, B.D., Kuncha, M., Kumar, J.M., Sudhakar, G., and Sistla, R., 2015. Oral administration of geraniol ameliorates acute experimental murine colitis by inhibiting pro-inflammatory cytokines and NF-κB signaling. ***Food & function 6***, 9, 2984-2995.

[22] Nikolaus, S., Schulte, B., Al-Massad, N., Thieme, F., Schulte, D.M., Bethge, J., Rehman, A., Tran, F., Aden, K., and Häsler, R., 2017. Increased tryptophan metabolism is associated with activity of inflammatory bowel diseases. ***Gastroenterology 153***, 6, 1504-1516. e1502.

[23] Pascal, V., Pozuelo, M., Borruel, N., Casellas, F., Campos, D., Santiago, A., Martinez, X., Varela, E., Sarrabayrouse, G., and Machiels, K., 2017. A microbial signature for Crohn's disease. ***Gut***, gutjnl-2016-313235.

[24] Paulson, J.N., Stine, O.C., Bravo, H.C., and Pop, M., 2013. Differential abundance analysis for microbial marker-gene surveys. ***Nature methods 10***, 12, 1200.

[25] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V., 2011. Scikit-learn: Machine learning in Python. ***Journal of machine learning research 12***, Oct, 2825-2830.

[26] Peng, H., Long, F., and Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. ***IEEE Transactions on pattern analysis and machine intelligence 27***, 8, 1226-1238.

[27] Robinson, M.D., Mccarthy, D.J., and Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. ***Bioinformatics 26***, 1, 139-140.

[28] Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C., 2011. Metagenomic biomarker discovery and explanation. ***Genome biology 12***, 6, R60.

[29] Sommer, F., Rühlemann, M.C., Bang, C., Höppner, M., Rehman, A., Kaleta, C., Schmitt-Kopplin, P., Dempfle, A., Weidinger, S., and Ellinghaus, E., 2017. Microbiomarkers in inflammatory bowel diseases: caveats come with caviar. ***Gut***, gutjnl-2016-313678.

[30] Stulberg, E., Fravel, D., Proctor, L.M., Murray, D.M., Lotempio, J., Chrisey, L., Garland, J., Goodwin, K., Graber, J., and Harris, M.C., 2016. An assessment of US microbiome research. ***Nature microbiology 1***, 1, 15015.

[31] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. ***Journal of the Royal Statistical Society. Series B (Methodological)***, 267-288.

[32] Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I., 2007. The human microbiome project. ***Nature 449***, 7164, 804.

[33] Vapnik, V., 2000. ***The nature of statistical learning theory***. Springer-Verlag New York, Inc., New York, NY, USA.

[34] Vázquez-Baeza, Y., Gonzalez, A., Xu, Z.Z., Washburne, A., Herfarth, H.H., Sartor, R.B., and Knight, R., 2017. Guiding longitudinal sampling in IBD cohorts. ***Gut***, gutjnl-2017-315352.

[35] Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., and Alm, E.J., 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ***The ISME journal 10***, 7, 1669.

[36] Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., and Birmingham, A., 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. ***Microbiome 5***, 1, 27.